REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
14-12-2005	Final Report	01-Dec-00 - 01-Dec-05

4. TITLE AND SUBTITLE

Mathematical Basis of Knowledge Discovery and Autonomous Intelligent Architectures-Voice Operated Flying Object

5a. CONTRACT NUMBER
ISTC Registration No: 1993p

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Andrey Ronzhin, Alexey Karpov, Izolda Lee, Yuri Kosarev

5d. PROJECT NUMBER

5d. TASK NUMBER

5e. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

St. Petersburg Institute For Informatics & Automation of the Russian Academy of Sciences
39, 14th Liniya
St. Petersburg 199178
Russia

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

EOARD
PSC 802 BOX 14
FPO 09499-0014

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
ISTC 00-7031-4

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Global awareness (GA) entails the acquisition of data from local to global levels, appropriate fusing of the data, and presentation of that data as useful information. This data will then be fused to fully describe situations of interest such as large transportation systems and complex communication systems. This project specifically aims at developing the mathematical basis, architecture and software techniques implementing particular new technologies to support Global Awareness and comprises six main tasks. Task 4 was:
4. Voice Operated Flying Object.

15. SUBJECT TERMS

EOARD, Mathematical And Computer Sciences, Computer Programming and Software

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18, NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			PAUL LOSIEWICZ, Ph. D.
UNCLAS	UNCLAS	UNCLAS	UL	134	**19b. TELEPHONE NUMBER** *(Include area code)* +44 20 7514 4474

Final report

ISTC-EOARD Project 1993P, task 4

VOICE OPERATED FLYING OBJECT

Authors:

Principal investigators of task 4: Andrey Ronzhin
Alexey Karpov
Izolda Lee
Yuri Kosarev

SPIIRAS, 39, 14th line, St-Petersburg,
199178, Russia
Tel: (812) 328-7081
Fax: (812) 328-4450
E-mail: ronzhin@iias.spb.su

November 2003
St. Petersburg

Contents

Introduction

In this Final Report the main results of the research of SPIIRAS Speech Informatics Group according to the Project #1993P, task 4 "Voice operated flying object" are presented. During the project all the topics of the technical schedule (Table 1) have been fulfilled.

Table 1. Technical schedule of the project 1993P, task 4

Activity or Event	Period of Project Activity
D-1. Autonomous modeling of speech interaction.	1-4 Quarters
D-1.1. Development of a dialog system.	1-2 Quarter
Technical Report to ISTC	1 Quarter
Interim Report #1	2 Quarter
Technical Report to ISTC	2 Quarter
D-1.2. Creation of a database of the autonomous system based on acoustical, situational and associative analysis.	3 Quarter
Technical Report to ISTC	3 Quarter
Interim Report #2, summarizing the efforts of the tasks D-1.1 and D-1.2.	4 Quarter
Demonstration of the DEMO-version of speech dialogue system.	4 Quarter
D-1.3. Autonomous system adjustment and testing. DEMO-version of the voice control system creation.	4 Quarter
Technical Report to ISTC	4 Quarter
D-2. Study of an aircraft (satellite) control specifics.	5-8 Quarters
D-2.1. Determining the forms of initial data required for modeling the voice control system.	5 Quarter
Technical Report to ISTC	5 Quarter
Interim Report #3	6 Quarter
D-2.2. Creation of automated database including inter-word associations and structure of situation arising on control of typical flying object.	6 Quarter
Technical Report to ISTC	6 Quarter
D-2.3. Creation a flexible data base structure allowing to take into account flying object specifics and get simply adapting to any new modification of the object work logic.	7-8 Quarters
Technical Report to ISTC	7 Quarter
Interim report#4- summarizing theoretical results concerning tasks D-2.1 - D-2.3 as a whole.	8 Quarter
Technical Report to ISTC	8 Quarter
Submission a paper in an International Journal	8 Quarter
D-3. Adaptation of a speech model to a flying object emulator.	9-12 Quarters
D-3.1. Compiling initial data for object control emulator. Creation of the model approximating to the logic of a flying object.	9-10 Quarters
Technical Report to ISTC	9 Quarter
Interim report#5	10 Quarter
Technical Report to ISTC	10 Quarter
D-3.2. Consideration of the flying object specifics for creating the	11 Quarter

Activity or Event	Period of Project Activity
database of the speech control model. Development of software, its adjustment and testing.	
Technical Report to ISTC	11 Quarter
D-3.3. Testing and checking the model. Preparing the final report.	12 Quarter
Final report - summarizing the results of the evaluation the voice control model concerning the task 4 as a whole will be delivered. The software code will be available well.	12 Quarter
Technical Report to ISTC	12 Quarter

Moreover the report contains the results of investigations, which were not initially included in the project schedule. These tasks are integral adaptation, continuous speech input and using extra-linguistic knowledge (situational analysis).

The first section briefly presents the main problems of human-machine interaction and the problems of speech dialogue such as adaptation, continuous speech and robustness. The typical model of man -machine interaction based on phrase understanding is considered. Also the specificity of the voice operated moving object is described.

In the second section the survey of the speech processing methods, which are used in the existent speech technologies, is presented. The main difficulties and problems of all the levels of the speech processing (speech detection, parametrical signal representation, isolated speech recognition, continuous speech recognition, high level language processing, speech understanding, using situational context [28,85,86,88,90,91,112,130,158,160,143]) are analyzed.

In the third section the approach of the Speech Informatics Group to speech understanding process is presented [144]. The mathematical model is based on the speech acts theory [5], the conception on integral processing [55,103] and also the results of psycho-physiological experiments on human speech perception [161].

The fourth section is devoted to theoretical research, which has been fulfilled in the framework of the project. In order to create robust and competitive model we researched all the levels of speech processing and carried out the wide spectrum of works.

In order to extract the speech signal from noise environments the method based on spectral entropy analysis was investigated and developed. Research of the parametrical representation of the speech signal was conducted during the whole period of the project. Two methods (sign autocorrelation function and spectral-difference features) robust to variations of the signal amplification level were developed [56,58,139]. Besides the second method has shown the robustness to accidental nonlinear spectrum deformations.

The problem of continuous speech recognition has existed over 20 years and this problem has not been solved yet. In the course of project we proposed the sliding analysis method of

continuous speech, which is robust to grammatical deviations in a pronounced phrase and has acceptable complexity for use in the speech understanding model [57,94,96,156].

We paid special attention to robustification of the speech dialogue model. It interests the researchers and consumers of such systems more and more now, because there were many unsuccessful attempts of real applications of speech technologies. In this section it is shown that the robustness of people speech perception is much better than that of the machine. Some examples of intuitive comparison of several well-known methods for speech understanding are adduced. An attempt is made to offer the definition of robustness degree within the framework of the integral conception, which is elaborated in the group [55,58].

Accuracy and flexibility of the system depend on its capability to adapt to various aspects of the concrete application. Moreover during the debugging and exploitation the system can obtain new data and so the databases adjustment is necessary. For this aim the procedure for integral adjustment of databases has been developed. The integral approach takes into account acoustic aspect, language aspect, subject area and also the integral optimization of the model parameters. It allows to achieve required efficiency of the databases adjustment and also the flexibility of the understanding model to new applied tasks.

As a result of the research the following modules were created: (1) the module of continuous speech recognition based on the sliding analysis of speech signal and a posteriori phrase hypothezation. This method is robust to grammatical deviations; (2) the integral adaptation module providing the inter-coordinated adjustment of all databases of the speech understanding model. The developed modules were included into the integral understanding model. As a result the united software complex has been created. It provides robust understanding of continuous speech as well as portability to new applied tasks due to the integral adaptation procedure [95].

In the conclusion of the fourth section the problem of extra linguistic information and its paramount importance in the speech understanding process is discussed. At present the approaches to formalization and use of situational information are poorly developed. Therefore the problem of using situational information leads to the problem of creating situational databases, which contain constraints of the applied area. In our works the model of the applied area for the restricted tasks connected with the control of technical objects (a car, a aircraft, a robot, etc.) is proposed [56]. But there are no techniques of the creation of situational databases for other intellectual applications. So the investigation of this problem presents interest within the framework of this project and its continuation.

In the fifth section the developed research prototype of speech understanding and the demonstration model of the voice operated flying object, which includes the aircraft emulator and the continuous speech understanding module, are described. The software classes of the speech recording, recognition and understanding are considered in detail. The manuals for operation with the research prototype, the demonstration model and additional programs for databases adjustment are adduced.

The sixth section is devoted to the most important directions of the research, which are required for creation of the speech technologies for mass usage. At this moment the group is beginning the research of speech data organization based on ontologies and in the future we are planning to use the multiagent system for effective collection and updating of speech and language information. Moreover there is the problem of using speech jointly with other modalities for creation of multimodal interfaces and their using in perspective intellectual systems of human-computer interaction.

In the seventh section the participation of the group in international conferences and the publications in the reviewed journals are described.

The conclusion of the report sums up the project. Besides, the theoretical results of the research, the developed model of the voice operated flying object are presented. Also the most important tasks and perspective directions of the research are discussed.

1. State of the art in speech recognition/understanding

Generally the concept of speech understanding consists of two stages: speech recognition (SR) and natural language understanding (NLU). This section briefly presents: (1) the typical structure of human-machine interaction in which a computer accomplishes the phrase *understanding* (interpretation); (2) the main problems of the speech dialogue; (3) problem state of speech understanding in different dialogue systems; (4) the specificity of the voice operated moving object.

1.1. Typical structure of human-machine interaction

The typical structure of human-machine interaction (dialogue model) is shown in Figure 1. A dialogue process is usually activated by a speech act. A human makes a decision about some act with the help of a priori knowledge about the theme and the aim of a dialogue, its language, and uses all the accessible current information about the dialogue progress, environment and current time.

The speech recognition module recognizes acoustic speech signal and produces the word sequence. This word sequence goes into the language understanding subsystem, which fulfills the meaning phrase representation by semantic frame, meaning type index, etc. During word sequence processing the understanding system uses the knowledge about the subject domain language and the current situation.

The dialogue management system (manager) accomplishes the coordination of all dialogue system components. In simplest systems this module is missing but with complication of dialogue process it is required for words check, prompting to operators, etc. Thus, the manager is an essential part of modern dialogue systems.

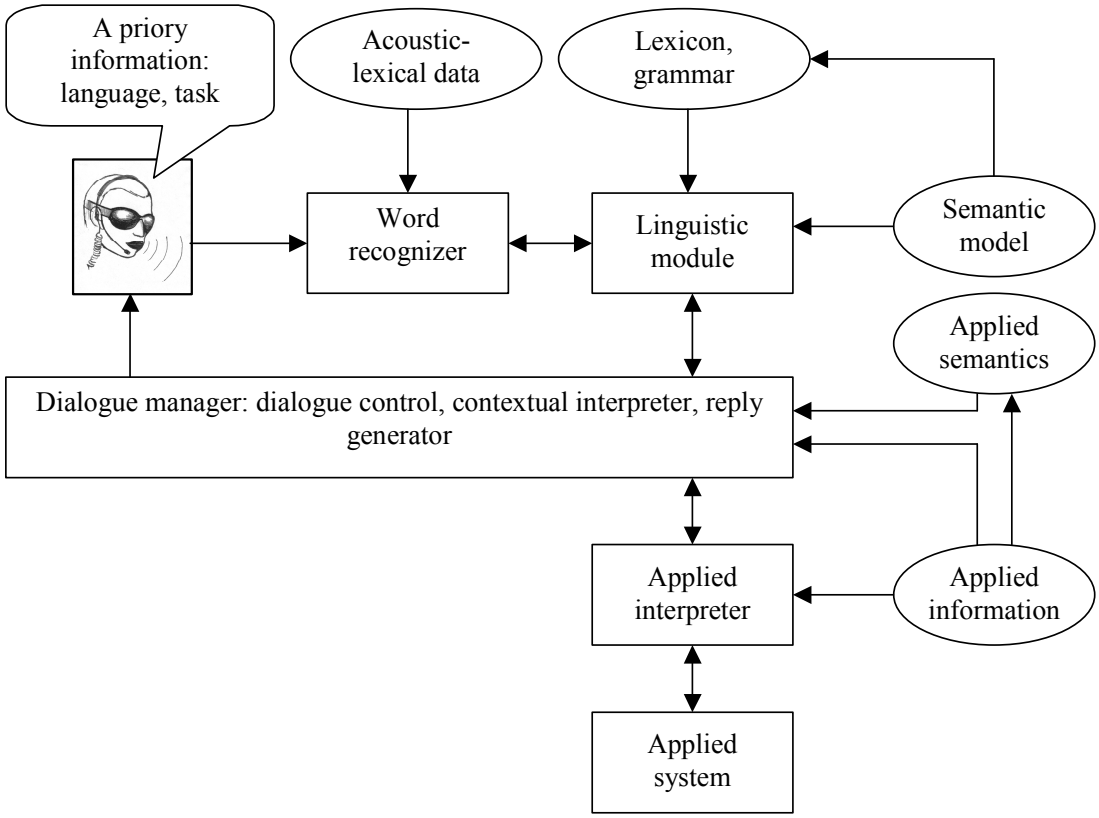


Figure 1. Typical structure of the dialogue model

It is necessary to notice that first dialogue systems used the word recognition and the menu system only [38,63] without the understanding level. Such approach cannot be named intelligible. Also the manager was not used or his functions were reduced to the confirmation of the speech recognition correctness by the user. The efficiency of such systems was insufficient.

At present the problem of dialogue management acquires important significance concerning the dialogue applications, which are rapidly developing and becoming more complex. Many research organizations deal with the investigation of dialogue strategies [31,35,108], dialogue structure [92,119], manager teaching [61,69,124] and others.

1.2. Main problems of the speech dialogue

The analysis of automatic systems for speech dialogue has shown that the most important issues are the understanding of the user's inquiries in the situational context of the object domain, spontaneous speech processing of an arbitrary user, adaptation to the acoustic environments and system adjustment to other aspects of the concrete application. Besides, numerous attempts to use speech technologies have discovered the problem of robustness, i.e. the system's capability to resist various casual influences. Thus, the main problems of the speech dialogue are adaptation, continuous speech processing and robustness.

1.2.1. The problem of the system's adaptation (to the user, the environment, the applied area)

Any model is more flexible and effective if it admits the possibility of the adjustment to the concrete conditions of the exploitation. This property is especially important when the application conditions are not constant and have some variations. In our case the speech recognition system must be adapted to three main factors: (1) the speaker, which interacts with the system by voice; (2) acoustic environment (noise, background, etc.), in which the dialogue between the system and the speaker takes place; (3) applied area, i.e. concepts, actions, intentions peculiar to the concrete object domain.

The most important property of the adaptive speech recognition system is the consideration of differences of the users' voices. These differences depend on the gender, the individual form of the vocal tract (which affects the pitch and the timbre of the voice), dialects, individual manner of sound and intonation formation, emotional state, etc. Besides, each speaker is characterized by the significant variation of pronunciation, i.e. the same human pronounces the same word with some difference. Also the physiological and psychological state of the speaker influences on the pronunciation quality essentially [161].

The problem of acoustic environment arises inevitably when using speech technologies for moving objects controlled by voice. Background noises (especially with unsteady characters) significantly degrease the recognition quality of the inputted speech signal. In order to reduce their influence differential microphones, a special noise shield helmet are used. Moreover the recorded signal is processed by filtration methods, which allow to separate the useful signal from the noise component. For this aim various adaptive filters based on spectral transformation are applied successfully [13,27,34,36].

Now the significant requirement to a speech recognition system is the capability to adapt to an applied area. In particular the language model can be adapted to the concrete dialogue specifics peculiar to this task [118]. Speaking about the control of technical devices by voice the changes of the control logic or object data require the adjustment of the mutual data by a human and a machine for effective interaction. Lack of adaptation or adjustment at this level makes the system unstable which leads to additional errors.

Types of adaptation mentioned above allow to make the performance of the speech dialogue system more robust to various deviations, which arise at the input of the system or during speech processing within the system. Thus, the integral approach to the adaptation problem is required, which takes into account the acoustic aspect, the language aspect and the applied area.

1.2.2. Continuous speech recognition problem

In contrast to the printed text or artificial signals natural speech does not allow simple and univocal dividing into elements (phonemes, words and phrases) because these elements do not have obvious physical boundaries [159,162]. They are segmented in a listener's mind, as a

result of the complex multilevel process of speech recognition and understanding. If we ask a listener to write down unknown foreign speech as a sequence of phonemes, he will make a lot of mistakes at the word and phrase levels. Therefore even a human cannot segment a speech without using the knowledge of lexicology, grammar and meaning [143].

Moreover the parametrical description of a word pronounced separately differs essentially from the same word pronounced in a phrase. It essentially complicates the word recognition in the stream of continuous speech. The existent approaches to continuous speech recognition are presented in detail below.

In order to provide the high robustness of the continuous speech understanding system to grammatical deviations, it is required that the most probable word chain be obtained at the recognizers' output independently of syntactical correctness of an input utterance. This task is easily solved by the isolated input, but the continuous speech input is more complex since the words of continuous speech do not have obvious physical boundaries.

Most of the methods use the *analysis through synthesis* principle, which leads to hypothesis generation methods when the word templates/models of the given vocabulary are combined for obtaining all possible hypothetical phrases. In case of the complete enumeration the complexity of such processing increases to an unacceptable amount. Therefore the developers began to use additional high-level information in the form of grammatical rules or stochastic N-gram. However, in this case the system is unable to process syntactically incorrect utterances, that degrades the systems' robustness to syntactical deviations.

Thus, the problem of creating the continuous speech recognizer without any high-level constraints exists. Therefore the original method for continuous speech recognition based on *sliding analysis* was elaborated. The results of the research are presented in Section 3.

1.2.3. The problem of robustness of the speech understanding process

According to the commonly accepted definition the robustness of language (text) processing is the ability to resist diverse grammatical deviations [99]. According to Selfridge's hypothesis robust speech understanding can be achieved with the integral approach implementation. *A robust understanding system is such a system, for which it can be guaranteed that it understands input utterances in spite of any word missing, any failure of word order, and availability of ellipsis both with and without correction. The system is based on semantic and syntactical knowledge of the object domain and the context.*

In automatic speech understanding syntactical errors/variations could be made by a speaker, or/and word recognition errors could be made by a system. Thus, to estimate the understanding robustness of a system, we have, at first to be able to estimate high-level deviations of inputted utterances by some quantitative measures.

For a long time the speech understanding modeling was based on sequential parsing ideas. In this conception the reliable model is presented as the composition of reliable components and any unreliable part sharply worsens the quality of the whole model. However, there are

many contrary examples in nature and technique, when the reliable whole is created from unreliable parts. For example, the problem of reliable data transfer by redundancy correcting codes was solved a long time ago: a message is decoded exactly in spite of a certain amount of wrong symbols. Speaking about natural analogies the synergism phenomenon should be mentioned, when the summary effect of organs or functions exceeds the simple sum of particular effects [136,152,155].

Below we compare the traditional model of consequential processing (Figure 2 left) with the integral model for automatic speech processing (Figure 2 right).

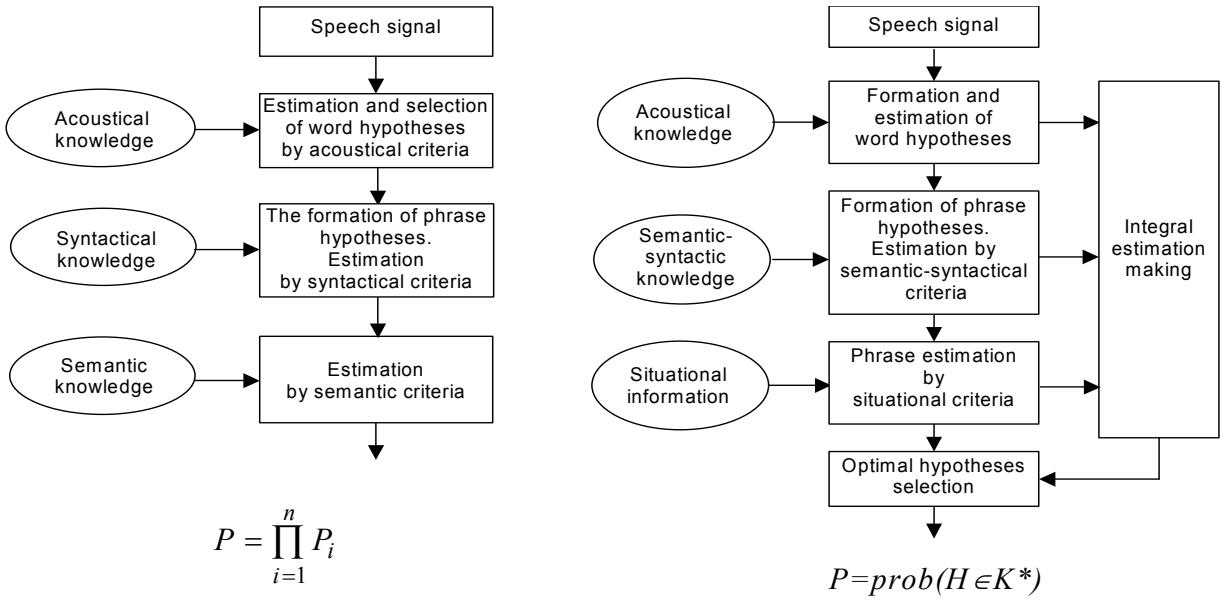


Figure 2. Sequential and integral models of speech processing

In the first model the search for «right units» goes at each level with transferring them to the next one. It is obvious, that the probability of successful passing through all the levels is: $P = \prod_{i=1}^{n} P_i$, where *n* is a number of levels, P_i is a probability of the hypothesis to be acceptable for the *i*-level criterion. This simple formula leads to the following conclusion: the more levels we use, the worse quality we get, i.e. the final accuracy cannot be more than the accuracy of any level.

Let us consider the typical speech understanding model known from many publications, which is based on sequential speech processing (Figure 3). It contains 3 base modules: the speech recognition module, the module of grammatical parsing and the dialogue manager. Such model has the following disadvantages: (1) the modules work consequently and it cannot principally give high understanding accuracy; (2) the syntactical/stochastic constraints of word recognizers do not pass the pronounced phrase with the partial incorrectness in the understanding level which leads to lowering the robustness to grammatical deviations at the input signal; (3) adaptability is provided at the acoustic level only. Moreover, different modules

and databases are created by diverse organizations and so the process of adaptation of the whole model to the concrete task is very complex and non-optimal.

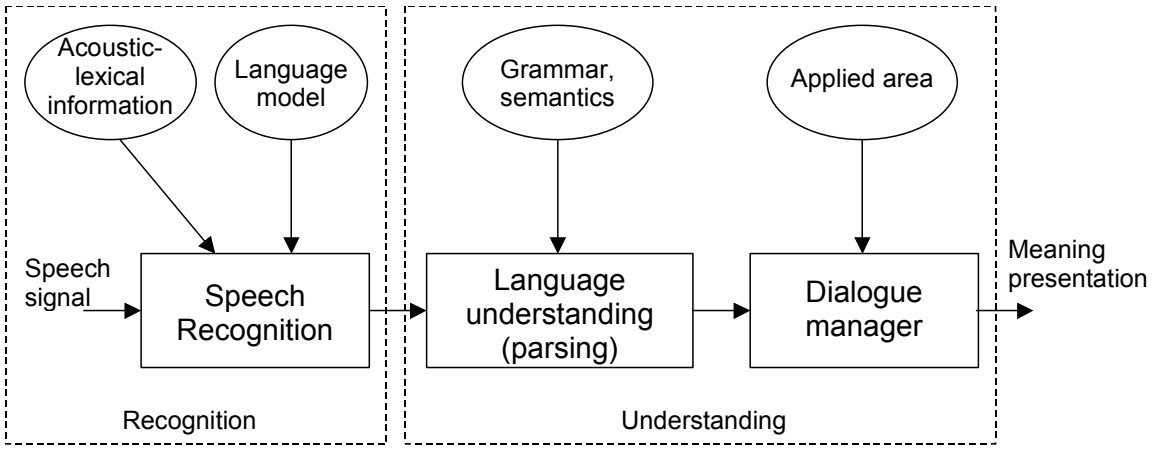


Figure 3. A typical speech understanding model

The matter goes otherwise for the integral speech processing. The influence of each level is reduced with the increase of their number. The probability of the right solution is the probability to be included in the corresponding semantic cluster K*: $P = prob\ (H \in K^*)$. This principle leads to more robust procedures, which well corresponds to the communicative theory, the speech psychology and our research [53,144]. Besides, the hypothesis of the integral processing adequately reflects the nature of a human's language processing. It leads to the assumption that integral processing produces robust understanding.

1.3. The specificity of the voice control for moving object

The system for the voice control of an object is an electronic system aimed for the transformation of the operator's (pilot, driver, etc.) voice commands into corresponding control signals performed by devices of the object. The system recognizes and understands an input signal using main kinds of information on the language, the operated object and the current task. The system also gives an operator all the necessary visual and voice information concerning the control process.

It is necessary to notice that voice control systems differ from all the types of automatic control systems. Since a human is a one who makes preparation and decision-making of the control acts. The task of the system is to correct understanding of human commands. However, it is possible to give deliberative *vote* to the system in order to solve vague situations.

An operated object can be a technical object, such as the means of transportation (an aircraft, a satellite, a ship, a car, etc.), a technological system, information processing, a robot, domestic devices, etc. For these objects the language of control (including professional languages), possible situations and logical-time relations in situations structure are usually studied well.

The examples of voice operated objects are shown in Figures 4 and 5.

Figure 4. Voice operated aircraft

The model for a human-machine interaction process is shown in Figure 6. The first generations of dialogue systems were based on linguistic automatons. Their application was to verify the truth of an input utterance (true or false). But nowadays the models, which allow to consider all the main units of the information process and the interaction between linguistic and extra linguistic information, are becoming more popular.

The amount of manual controls (buttons, switch keys, tumblers, etc) in modern technical objects increases permanently and becomes immense. So the voice control can be used as a worthy substitution of manual one to provide high reliability, safety and efficiency. Moreover, voice control is preferable in abnormal cases: extra mechanical overload, zero gravity, low visibility, work in heavy suits (a diving-suit, a space-suit), etc.

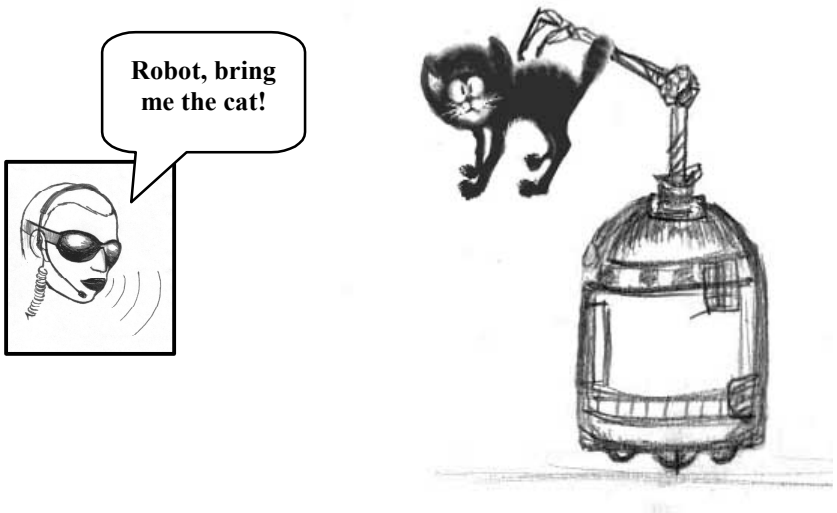


Figure 5. Voice operated robot

Moreover, the dialogue during the speech control of a moving object has some peculiar properties. There is strict structuring of movement and object functions, as well as the real-time factor. The determination of the object's functions puts the significant limitations on the

dialogue language. The rational use of the situational constraints can essentially increase efficiency and robustness of speech control.

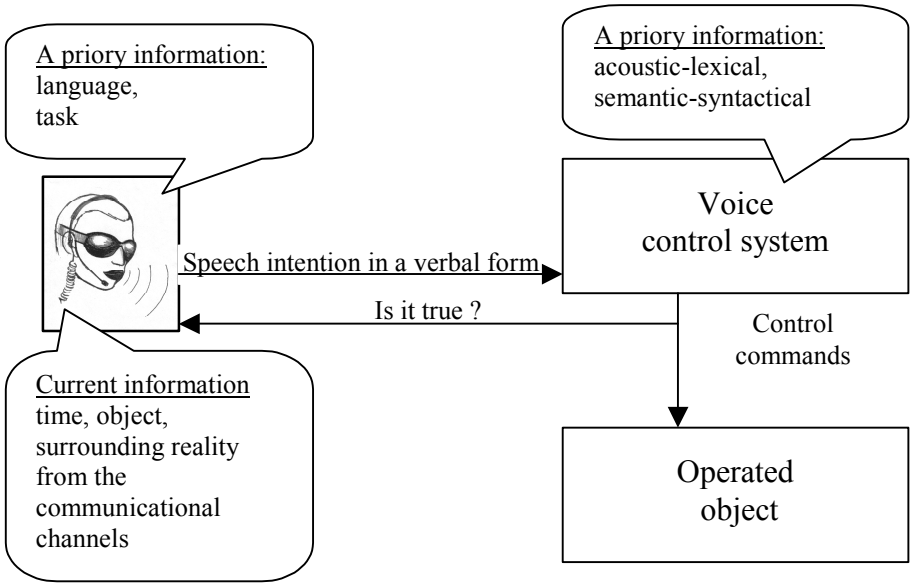


Figure 6. The model of human-machine interaction

In contrast to many other forms of the dialogue the real-time dialogue presents essential difficulties, because there is no time for repetitions, clarifications, etc. This difficulty may be overcome only owing to the analogies with the human's speech behavior of the dialogue partners in similar situations. The imitative modeling stage is necessary here. Two humans (the operator and the executor) take part at this stage. As a result all the necessary information for building of the dialogue language (sets of equivalent commands for all possible actions) and for the dialogue structure must be obtained. Also the information, which is required for the creation of a voice control system, must be obtained. Finally, such a communicative style must be produced, which allows to reduce the number of repetitions and errors to the minimum.

2. Methods for speech processing for voice control task

The process of speech recognition is the transformation of an acoustic signal captured by a microphone into a word chain.

Speech recognition is characterized by many parameters such as environmental noise, features of the speech communication channel, vocabulary size, speech variation, speaking mode (isolated words or continuous speech). Recognition of isolated words requires to make brief pauses between words. It slows down speech entry and decreases the naturalness, whereas the continuous speech input does not require such constraints (of course having its own difficulties).

The difficulty of speech recognition problems is mainly associated with the variation of speech parameters, on which many factors effect. At first, the speech production is a stochastic

process that leads to description variety of the same word pronounced by the same speaker. Moreover, the same signal can be divided into different number of words because of the ambiguity of word boundaries. The important factor is individual differences of vocal mechanism of various speakers. It is necessary to note the influence of the speaker's sex, age distinctions, dialect, emotional and physiological state of a speaker. The influence of the acoustic aspect (the microphone change, the microphone position relative to the mouth, acoustic conditions in a room, etc.) is significant too.

In order to solve the "speaker-dependence" problem it is important to select speaker-independent parameters at the level of parametric signal representation. Furthermore it is possible to use adaptation methods, which help to adjust the system parameters to the specific voice and acoustic environment.

The quality of acoustic-lexical word recognition is mainly estimated by the recognition accuracy of the concrete vocabulary. The possibilities of this level are evenly constrained by foregoing factors and the accuracy even for a man can be, for instance, only 85% at the 300-word vocabulary [77]. In this experiment syntactic-semantic relations were excluded from inputted speech and a human tried to recognize words using just acoustic-lexical information.

The word recognition rate is essentially decreased with increasing the vocabulary size, as the groups of acoustically similarly words are appeared which leads to confusion. There are two ways to solve this difficulty. The first is to use the sufficient hypotheses set, which surely contains really inputted words from these hypotheses. Here the principal possibility of recognizing right meaning is saved by an understanding system. However, the hypotheses set increases fast with the growth of the vocabulary size and for the vocabulary of 50000 words the set contains thousands of hypotheses at every inputted word.

The other approach is based on the best word hypotheses selection with using some high-level information. Syntactic and semantic information is usually used here as a stochastic language model. One of the first works in this area was IBM recognition model of 5000 words based on the combination of acoustic-lexical processing and the language model (bigrams and trigrams). The recognition accuracy of this system was 95% [45]. Such paradigm is used at present too. The general direction in acoustic processing is the development of techniques of hidden markov models (HMM), which provide the speaker independence and continuous speech recognition. At present there are several commercial systems with vocabularies of 60000 words and more [7].

However, it is necessary to note that such models were initially created for automatic stenography systems, which presuppose the subsequent manual correction. At present these systems are used in different intelligent applications such as information systems [63], voice control [30], translation systems [113,122], etc. However, these attempts don't achieve desired results because of incorrect usage of high-level information. Besides the understanding level is missing in such models, but observations show that these processes are present in the work of an interpreter, a stenographer, an editor, etc.

These disadvantages have been taken into account during the development of the integral understanding model.

2.1. Speech endpoint detection

In order to begin the speech recognition process it is required to detect the boundaries of speech in context of background noise. It is obvious that the accuracy of recognition of isolated words depends sufficiently on the accuracy of the detection of words boundaries. For continuous speech it is necessary to find the moments of the beginning and ending of the pronounced phrase too. The complexity of speech endpoint detection is connected with the peculiarities of the pronunciation of the concrete speaker, the presence of noise in his articulation process (aspiration, lip smacks, etc). Besides, the consideration of outside noises is very important, too. Some methods used for speech endpoint detection are presented below:

1. The Voice Active Detector (VAD), which is used in mobile phones, is based on differences in spectral characteristics between speech and noise.
2. The speech endpoint detector proposed by L. Rabiner and his colleagues uses the signal energy and the number of zero-crossings.
3. The original approach proposed by Auditech company.
4. The multi-channel algorithm for speech segment endpoint detection proposed by I. Mazurenko.

1) The Voice Active Detector (VAD) is used in GSM telecommunication networks. It detects time intervals when a user speaks [26,105]. For the fulfillment of the required functions the VAD must have a quick reaction that not to lose words and not to miss the useless fragments of silence in the end of a sentence. The VAD must work correctly in spite of the influence of background noise.

The VAD evaluates the energy of an input signal and, if it exceeds a concrete threshold, the detector activates data transmission. If the detector rejects all information until the energy does not reach the threshold then the initial part of speech activity will be truncated. Therefore the realizations of VAD demand to save in memory several milliseconds of information, in order to have the possibility of starting data transmission before the beginning of a speech activity period.

In standard GSM the VAD uses the spectral processing. The structural scheme of VAD is presented in Figure 7. The work of the detector is based on differences in spectral characteristics between speech and noise. It is assumed that background noise is stationary during a long time and its spectrum is changed slowly too. VAD detects the spectral deviations between the input signal and background noise. The inverse filter, the coefficients of which are calculated in the presence of background noise only, makes this action. In case of existence of speech and noise the inverse filter suppresses the components of noise and integrally decreases its intensity. Energy of the mixture “signal+noise” at the output of the inverse filter is compared with the threshold, which is calculated in the presence of noise at the input of VAD. This

threshold is higher than the level of noise energy. The excess of the threshold means the presence of mixture “signal+noise” at the input of VAD. The coefficients of the inverse filter and the threshold value are changed according to the current value of the noise level at the input of VAD. Since these parameters (coefficients and the threshold) are used in VAD for speech endpoint detection then VAD cannot decide when they must be changed. This decision is made by the second VAD based on comparison of spectrum contours in consecutive time moments. If they are analogous during long time then it is assumed, that pure noise is present and coefficients and the noise threshold can be changed (i.e. adapted to the current level and spectral characteristics of the input noise).

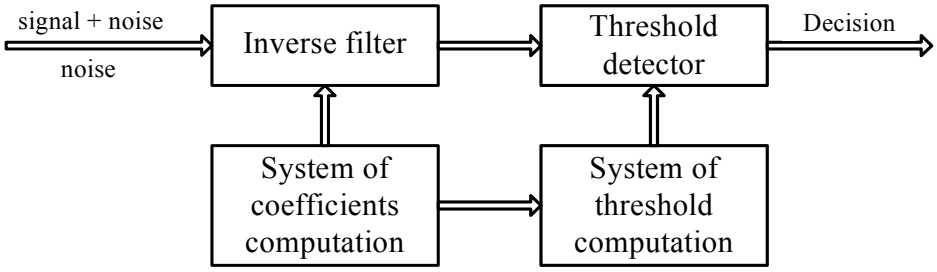


Figure 7. Structural scheme of the Voice Active Detector

2) Figure 8 shows the speech detector proposed by L. Rabiner and his colleagues [91], which uses the values of short-time energy of signal for speech endpoint detection. In this approach the adaptive level quantization module estimates the level of the acoustic background and uses the result to equalize the measured energy contour. Preliminary energy pulses, which are speech like bursts of energy during the recording interval, are then detected from the equalized energy contour. Finally, these potential energy pulse endpoints are ordered, according to their likelihood, to determine the possible sets of word endpoint pairs. Extensive experimentation is necessary in deciding the best values for the required set of thresholds and the ordering logic to provide the most reasonable sets of endpoints.

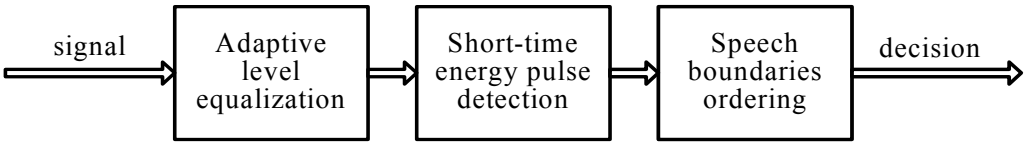


Figure 8. Algorithm for speech endpoint detection

3) The method based on the combination of three-edged filtration and dual-threshold procedure of the search of the word beginning/ending moments with iterative refinement of results [133], was proposed by the Russian company Auditech. The essence of this algorithm is the following. For each of three proportional spectral bands the points, where the signal begins to exceed the threshold (calculated out of the average level of noise in the channel), and the points, where the amplitude is lower than the threshold (calculated out of the average value of

the signal in the supposed word), are searched. Then the logical processing of mutual location of these points is used for detection of speech and silence.

4) The multi-channel algorithm for speech segment endpoint detection was proposed by I. Mazurenko [147]. This original approach uses the additional sensors (besides a traditional microphone) such as: additional microphones, a photosensor, a low-frequency sensor of sound pressure, the sensor of air stream. The essence of the multi-channel algorithm is described below. Some numerical parameters (signal energy, derivative of energy, etc.) are calculated using the information from the outputs of sensors (channels). The presence (or absence) of the speech signal in a concrete channel can be detected by comparison of these parameters with the fixed threshold. The time interval, where the signal exceeds the threshold in each point, is called the "quasi-speech interval", but the time interval, where the threshold exceeds the signal, is called the "pause". The boundaries of speech segments in each channel are calculated using a priori parameters such as: the maximal duration of the "pause" inside a speech segment, the minimal duration of the maximal "quasi-speech interval" inside the speech segment, minimal and maximal durations of the speech signal, etc. It is assumed that the fact of the speech segment is determined if there is a time moment, when the "quasi-speech interval" is simultaneously discovered in all channels. The boundaries of speech are determined by means of fusion of all "quasi-speech intervals".

Below the developed method for speech endpoint detection based on the calculation of entropy of spectrum will be described. This method allows selecting the speech (with required efficiency) in conditions of non-stationary noise with high amplitude.

2.2. The methods for parametrical signal representation

During creation of a speech recognition system the important task is the selection of speech parameters, which can sufficiently well differentiate the sounds and words of speech and at the same time be invariant to the particularities of the concrete speaker, change of acoustic environment, change of a microphone, etc.

There are many methods for parametrical signal representation. First speech recognition systems used autocorrelation analysis, hardware band-pass filtering, computation of signal spectrum, the method of linear predictive coefficients (LPC). At present two approaches are used: spectrum analysis and LPC. These approaches are very popular as they are well coordinated with models of hearing perception and speech production, respectively.

Speech waveform is usually digitized with frequency 8 - 22 kHz. This digital samples sequence is divided into speech segments with duration 10 – 20 ms. The feature vector is computed for each such segment. These vectors are the base for parametrical speech description. Such analysis is called short-time analysis.

The digital filter bank is one of the most fundamental conceptions in speech recognition. A filter bank corresponds to the model of human hearing. The spectral signal description received by the filter bank can be considered as the reaction of a mechanical system of internal and

middle ears by the influence of the complex signal. Here diverse components of this fluctuation force the vibration of separate parts of a basilar membrane [84].

The algorithms for feature extraction use preliminary processing of digital samples, for instance, the preemphasis filter and the weight window procedure [75,88], Fast Fourier Transformation, spaced equidistantly or by the specific nonlinear law according to *Mel* [23] or *Bark* scales [126]. Then FFT samples included in each filter are reestimated by the triangular window. The integral energy is computed and the logarithm of each filter output is calculated. Then the cosine transformation is applied to this data set. As a result the cepstral coefficients are obtained [91,106]. The cepstral coefficients obtained using Mel-scale are known as Mel Frequency Cepstral Coefficients (MFCC).

Linear Prediction Technique [75], based on autoregression analysis, is less popular today in speech recognition but it is still widely used in compression speech systems. The main principle of linear prediction is based on the prediction of signal samples by means of linear combination of neighbor samples. Prediction coefficients are weight coefficients used in linear combination, which are computed by minimization of average squared difference between the speech signal samples and their predicted values. Figure 9 shows the algorithm of the computation of features vectors by LPC. As a result we obtain the description, which is approximately identical to the vocal tract model. More complete speech description is received using the combination of the information about the vocal tract and the excitation source.

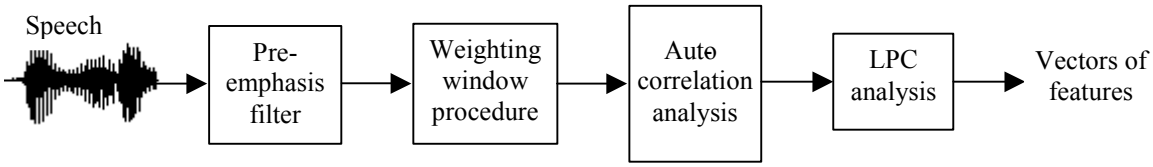


Figure 9. Speech processing by LPC

As the storage of parametric speech representation requires too much memory and taking into account the significant stochastic component of speech it is reasonable to compress this information. For this aim the vector quantization technique is used [74]. The application of this technique is presented in Figure 10. The base idea of this technique is the reflection of the unconstrained set of features vectors in the limited number of typical acoustic states. The vector of multivariate parameters is transformed into the index of the corresponding element of a codebook.

The important questions of the vector quantization are: creation of optimal codebook required to quantization, choosing the distance measure between vectors and the procedure of the search of the nearest element. For this aim the comparison with all elements of codebook or hierarchical clustering procedure are used [48,49,50,141]. The process of the codebook building is known as training process or the codebook filling. The most popular methods are *K-means* [72] and ISODATA [8].

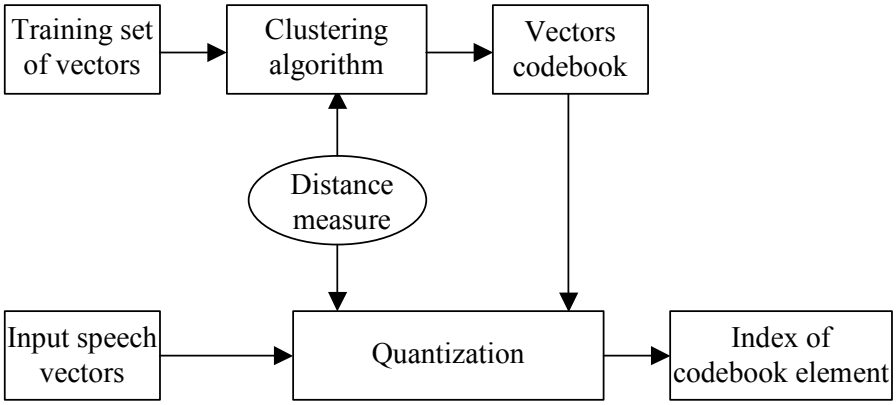


Figure 10. Scheme of the vector quantization technique

2.3. Isolated speech recognition

The key question in speech recognition is the identification of speech elements (phonemes, words, words sequence), which is performed by the technique of pattern comparison or by likelihood estimation of the signal correspondence to some class. These methods are considered in detail below.

2.3.1. Speech recognition by dynamic programming

The main difficulty in comparison of speech segments is connected with variations of time scale and its nonlinear fluctuations. So these algorithms are closely connected with the task of neutralization of time warping. The variability of speech temp is expressed in uncontrolled fluctuations of duration of speech phones, their segments and pauses. In order to compare a word with a pattern the matching of segments, corresponding to the same phonemes, is required. Then residual differences between the segments are computed and summed up using some weight coefficients.

Dynamic programming algorithms (DP) were commonly used for nonlinear speech matching. These algorithms are based on fundamental works of Bellman [12]. One of the first publications about using DP for speech recognition belongs to T. Vintsyuk [114]. NEC specialists studied this method in detail [97,98].

The objective of DP task is the search of the optimal matching for two speech parts *A* and *B*. Let the descriptions of words be presented as sequences of features vectors (Figure 11):

$$A = \{a_1, a_2, \ldots a_j, \ldots a_k\};$$
$$B = \{b_1, b_2, \ldots b_j, \ldots b_n\}.$$

Some measure *d(i, j)* is selected in features space, which allows to define difference degree between vectors a_i and b_j. Then the optimal path is searched. DP algorithm is based on using the recurrent equations. Recurrent DP equation, which allows work with two-time degree of warping is presented in Figure 12.

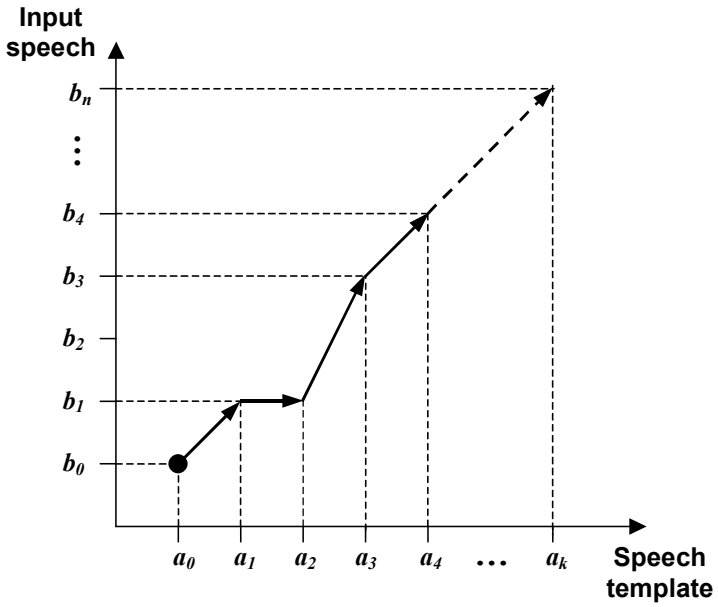
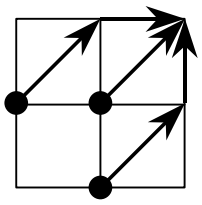


Figure 11. Matching of speech parts by DP method

The modifications of these equations are connected with the hypothesis about mutual warping character of speech segments. The study of this question [97, 143] showed that the optimal warping is two-time one. It was confirmed by the minimum of recognition errors.



$$g(i, j) = \min \begin{cases} g(i-1, j-2) + 2d(i, j-1) + d(i, j), \\ g(i-1, j-1) + 2d(i, j), \\ g(i-2, j-1) + 2d(i-1, j) + d(i, j). \end{cases}$$

Figure 12. Recurrent DP-equation with two-time degree of warping

Figure 13 presents the speech recognition model based on DP, which contains typical blocks: preliminary signal processing by LPC or spectrum analysis, boundaries detection of speech, comparison of input speech with speech templates from vocabulary. The result of DP matching is DP-distance, which allows to define the optimal hypothesis of the pronounced word.

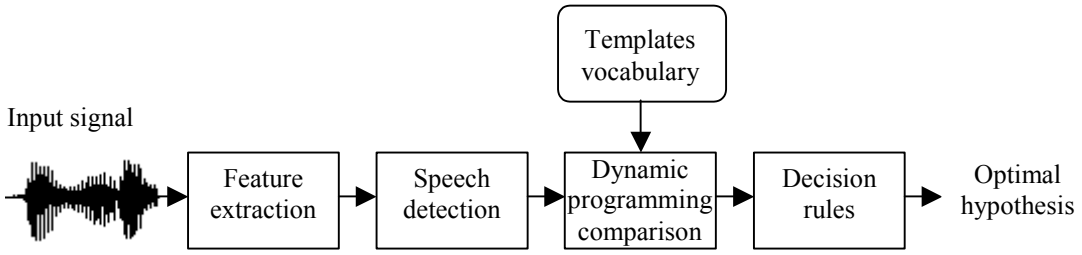


Figure 13. Speech recognition model by dynamic programming

Speech recognition by pattern comparison technique has been losing its significance recently in connection with the fast development of HMM methods. However, some modifications of DP methods are successfully used for continuous speech recognition.

2.3.2. Usage of Hidden Markov Models in speech recognition

The theoretical base of statistical Markov modeling was established by St. Petersburg's professor A. Markov at the beginning of the XX-th century. Today Markov modeling methods are used almost by all speech researchers in the world [46,85,91,149].

The Markov model modification which is called the Hidden Markov Model (HMM), is based on the theory of discrete random chains. It was firstly introduced and studied in the late 1960s and early 1970s [44]. HMM is a double stochastic process with one stochastic process, which cannot be directly observed (it is hidden), otherwise than through the other stochastic process, which produces the observations sequence [64]. The models of such type are especially suitable for the speech signal description.

Markov models have efficiently informative mathematical structure, so they have become the theoretical basis in different research areas (not only language research). HMM can improve the quality of the signal mixed with noises and corruptions by means of modeling the speech signal source, dialogue structure optimization, etc.

For building the recognition model based on HMM the following main parameters are chosen: model type (ergodic, Bakis model [9,110], etc), model size (number of states) and the type of observed parameters. The speech recognition unit may be a phoneme (or phoneme-like units), diphones, semisyllables and syllables [6] and also produced units such as phenems, phenons and acoustic segments [115]. Every such unit may be presented as some HMM, for which the estimation of parameters is performed by speech training data set. The transitions structure for a phonetic unit according to Bakis model is shown in Figure 14.

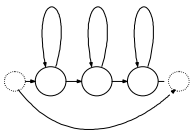


Figure 14. The model of a phonetic unit

The HMM of word is created by concatenation of the models of phonemes (from an alphabet) that is shown in Figure 15.

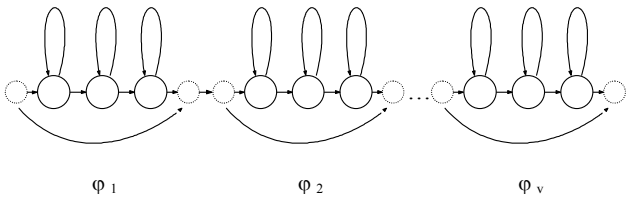


Figure 15. Word model composed by phonemes from alphabet φ

The essence of HMM speech recognition is likelihood estimation that the input sequence (observed) corresponds to the hypothesis model [64,98,115]. The using the Markov modeling for isolated words recognition may be divided into two stages. HMM building for every word of the vocabulary with the size V and also the optimization of their parameters by the training process. The second stage is word recognition. This processing is shown in Figure 16, where $O = \{o_1, o_2, \ldots o_n\}$ the sequence of the observations is defined by the signal analysis. The calculation of models likelihoods is accomplished for all possible hypotheses $P(O/\lambda^v)$, $v \in [1, V]$. The model with the highest likelihood is the optimal hypothesis of the pronounced word. So the index of the recognized word v^* is computed as:

$$v^* = \arg\max_{v \in [1,V]} [P(O/\lambda^v)]$$

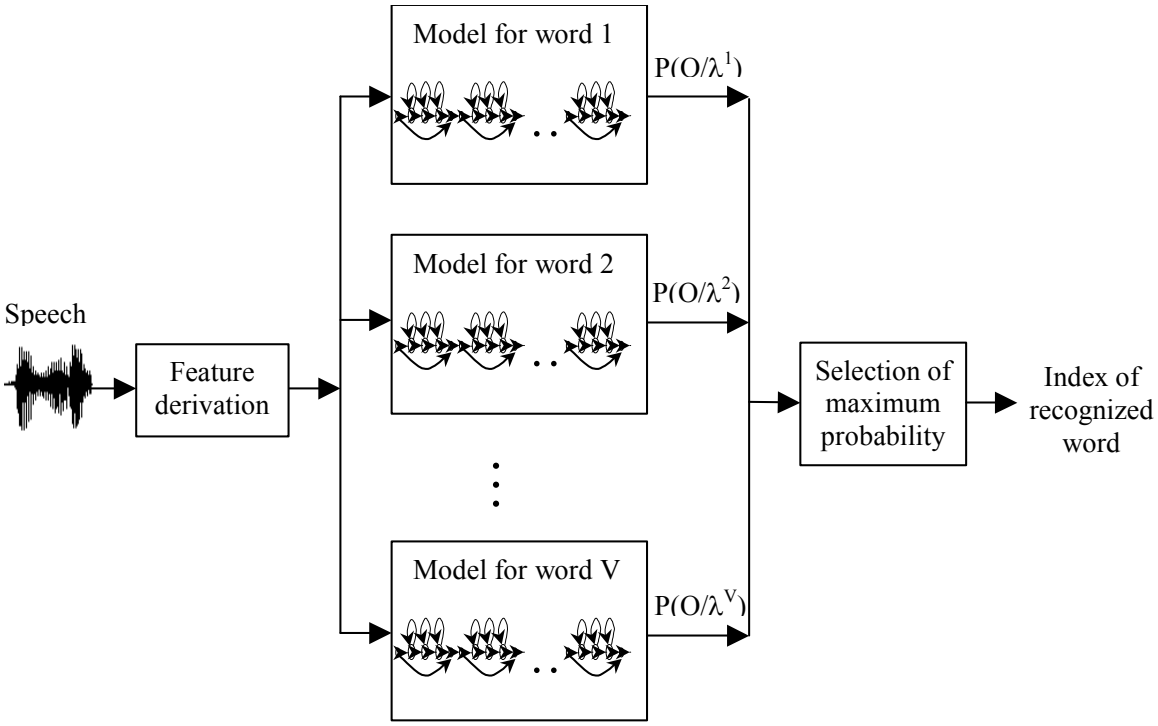


Figure 16. Isolated word recognition based on HMM

The development of the recognition systems based on HMM requires significant amount of acoustic data. Nowadays there are many acoustic corpora [40,41]. During the process of the creation of speech databases many factors must be taken into account such as speaker's characteristics (nationality, sex, age), the data-communication channel (microphone, telephone), the noise level. These databases contain phonetic transcriptions, the labeling of an acoustic signal into phonemes, syllables, words and phrases.

The past decade saw a significant progress in the speech recognition technology based on HMM. Substantial progress has been made in the basic technology which has led to lowering of the barriers of the speaker independence, continuous speech and large vocabularies.

2.4. Approaches to continuous speech recognition

In contrast to the printed text or artificial signals natural speech does not allow simple and univocal dividing into elements (phonemes, words and phrases) because these elements do not have obvious physical boundaries. They are segmented in a listener's mind during the complex multilevel process of speech recognition and understanding. If we ask a listener to write down unknown foreign speech as a phoneme sequence, he will make a lot of mistakes in words and phrases. Therefore even a man cannot segment speech without using the knowledge from lexicology, grammar and meaning.

Words boundaries in continuous speech cannot be detected without using total sum of a priori knowledge of the language and its specific application. The boundaries can be detected just in the speech recognition process by selecting the optimal word sequence, which is best of all corresponds to an input speech stream according to acoustic, linguistic and pragmatic criteria. Since the optimal selection is performed simultaneously with speech time normalization in conditions of words boundaries ambiguity so the multilevel optimization must be used. There are three popular approaches to solving the continuous speech recognition problem: the two level dynamic programming method, the level building method and the time synchronous level building (one-pass) method [91]. The algorithms use similar base principles and differ in their computational efficiency, storage requirements and simplicity of realization in real-time hardware.

These methods are used both by speech recognition based on pattern comparison (DP methods) and by statistical modeling based on HMM. In the latter case *N*-state HMM is used as speech pattern presented as *N*-frame vectors sequence. The speech recognition task consists in selecting the optimal sequence of word models, which have maximum likelihood to an unknown input word sequence.

The important part of continuous speech recognition is the model of generation of the phrase hypotheses (language model), which are usually based on grammar or stochastic constraints.

One of the most popular statistic models is N-gram model. The aim of the statistic model is the estimation of probability of occurrence of the word chain $W=w_1w_2...w_q$ in recognizable signal. Using probability $P(W)=P(w_1)P(w_2|w_1)...P(w_q|w_1...w_{q-1})$ we must calculate $P(w_j|w_1...w_{j-1})$ for each j. The probability of occurrence of each word is calculated as the function from previous $N-1$ words. Nowadays practically all systems of speech recognition use N-gram model. The probability of the whole sentence is calculated as the production of its N-grams probabilities. The simplest model is 1-gram binary model. Where the probability $P(w_j|w_k)=1$, if such word sequence is possible in a language and 0 otherwise. In general case $P(w_j|w_1...w_{j-1})$ is estimated as the relation of sequences $w_1...w_j$ to $w_1...w_{j-1}$ on the training set. However, in practice training data is always incomplete and a part of theoretically possible n-grams is absent or occurs rarely to apply statistical methods for estimating probability of their appearance. If

such n-gram is found during the work then the correct hypothesis will be rejected or its probability will be very low. At that the probability of occurrence of such n-grams is corrected taking into consideration other n-grams. For example the trigram can be estimated as: $P(w_3|w_1,w_2)=p_1*F(w_3|w_1,w_2)/ F(w_1,w_2)+p_2* F(w_1,w_2)/ F(w_1) +p_3* F(w_1)/N$, where N is size of the corpus; $p_1+p_2+p_3=1$ and $p_1,p_2,p_3=>0$.

The main advantage of present models is the possibility of building a model by the training corpus with the large size and high speed of work. The main deficiency is the incorrect assumption about the independence of probability of a word from more longer history, that makes difficult modeling deep relations in the language.

Grammar is a system of rigidly fixed rules describing correct language sentences. Usually researcher makes them manually which is connected with some difficulties. But the quality of such grammars is significantly better than for n-gram models. Unfortunately, these language models have some deficiencies. They are too rigid and do not pass the sentences with any grammatical inaccuracies.

Also decision trees are used as language models for estimating the probability distribution for the next word from its history. A decision tree is a binary tree, in which each leaf is a probability distribution in the vocabulary and other nodes are predicates defined on the history set. As the result the path from the root to one of its leaves is obtained.

In practice stochastic or generative grammars with rigid syntax are usually used. However, in this case the model practically does not perceive syntactically inaccurate but potentially understandable phrases that makes difficult to define the correct meaning. The best model could be the model of the generation of all combinations of words as it is done for the recognition of continuously pronounced digits. However, in large vocabularies the number of possible hypotheses is too large. Therefore the corresponding methods for optimization are needed here.

2.5. Natural speech understanding

In this Section the necessity of the understanding level is discussed and a brief review of the existing methods for speech and text understanding are presented. Also we give the arguments for the integral paradigm for speech processing in comparison with well-known approaches.

2.5.1. Speech understanding as a necessary part of speech processing

What for is speech understanding required? This question did not disturb the first speech researchers because the vocabulary size was small (dozens of words) and one-word commands were used only.

However, this problem has been disturbing philosophers for a long time. The understanding problem was important already in ancient world concerning diverse kinds of controversy (in politics, in commerce, in lawsuit, etc.) when a necessity arose to interpret a crime in accordance with current laws or morals.

But a real boom arose in 1950s in the framework of the discussion "is a computer able to think?" In that period different approaches arose for high-level text processing and later – for speech processing.

The main aspect of understanding process is meaning interpretation, i.e. the transformation of a word chain into the certain meaning representation. The difficulty of this process is concerned with a lot of different distortions contained in the pronounced phrase.

In connection with the appearance of large vocabulary recognizers the wish appears to implement them in different intellectual applications. But researchers faced the robust speech understanding problem and the knowledge integration problem. It turned out that large vocabulary recognizers cannot decide the speech dialogue problem.

Thus the robust speech understanding process is the main problem for the speech dialogue research.

2.5.2. Speech understanding methods: two base paradigms

In dialogue systems the understanding problem is the translation of a word sequence obtained from the output of words recognizer to some meaning representation. The complexity is connected with great ambiguity of the recognized words sequence. The causes of distortions may be word recognition errors, acoustic noise, background conditions, heavy breathing, aspiration, lips smacks, cough, sounded pauses (…eee.., …mmm.., etc), repetitions, interruptions, word omissions, extra word insertion etc. Thus, the base problem of speech understanding is the neutralization of the described above deviations.

In according to international publications the problem of spoken language understanding (SLU) can be decided by two diverse paradigms: consecutive analysis and meaning recognition by quantitative methods.

The first paradigm is realized in the linguistic theories by means of Fillmore's theory of semantic cases, the conceptual dependencies theory, the semantic networks theory, etc. They are intended for *the meaning calculus* from the input words based on certain rules. In other words it is based on the presumption that language is a logical object and the thought is logical calculus. The models based on this approach have some disadvantages. Firstly, the adherents of this approach go deeply into to the details of the propositional calculus process to the detriment of the communicative goals of the speech dialogue. Secondly, as a rule, the propositional calculus is based on parsing and each mistake leads to the wrong decision. This approach does not consider the language variability, its ambiguity and context dependency.

In practice the first paradigm is realized by methods of consecutive grammatical and semantic analysis [14,42,117]. The result of this analysis in most cases is the filled structure of semantic frame [3]. Unfortunately the systems built on these principles usually achieve low accuracy of meaning recognition (about 70%) and they cannot be implemented in commercial applications. To overcome this disadvantage in modern systems the stochastic component in grammar [15,100] was implemented in combination with stochastic language model [25,87].

These systems contain a parsing model and a stochastic language model (LM). Further improvement of these methods is the addition of information about specific applied task where additionally to the usual language model the situational language model (LMS) [65] is used, which is analogous to a usual LM but deals with the specific applied task. The adaptation of such system to a task increases the dialogue's stability.

Concerning the paradigm of consecutive parsing it should be noticed that idioms, utterances of professional slang, all possible reducing, etc satiate the natural speech. Consequently it rarely allows filling all frame slots. Therefore it requires reentry of some word, entry of qualifying words that leads to the dialogue delay and does not correspond to the psychology of human conversation.

The essential disadvantage of the consecutive analysis paradigm is that semantic analysis requires preliminary grammatical analysis that leads to the problem of the errors detection and correction [60,121]. To decide this problem the methods for mistaken phrases accumulating to creation of enormous database of erroneous utterances could be used. Such disadvantage cannot be found in the systems based on the second paradigm.

The second paradigm appeared in contrast to the first one. This paradigm supposes that the main goal of speech communication is a speech intention transfer and the main goal of a listener is the recognition of this intention (but not a true/false estimation). This paradigm is well-coordinated with the speech acts theory and it becomes more popular.

The "Speech Acts Theory" (SAT) proposed by J. L. Austin, J. R. Searle and others [5] is one of the important linguistic theories, which made significant contribution into the speech processes studies. A speech act, which is pronounced in the direct contact taking into account the situation context is a subject of investigations in SAT. SAT provides more adequate understanding of the essence of the dialogue and that is very important for the creation of a dialogue system. Speech act is considered as intention to achieve the concrete desired aim.

Besides, this paradigm allows to use analogies from the communication theory i.e. the well-known principle of the redundant messages decryption by "minimum deviation".

At present there exist meaning recognition methods based on quantitative analysis (stochastic processing) and quantitative comparison of a word sequence with sentences templates contained in the databases [18,47,62]. The comparison is produced on the base of stochastic rules [62], heuristics [53,54] or using HMM modifications [38]. Unlike the first paradigm such system is capable of producing the correct response even in conditions of incomplete information, reduced phrases, particular distortions, etc. Such approach is generally more operative and, in our opinion, it is preferable for voice control tasks and other intellectual applications.

2.6. Using the situational context in the high-level speech processing

The natural language used in dialogues depends on the speaker's intention. The intention itself does not have exact formulation, but it puts the hidden guiding lines, which help to constrain the ambiguity of utterances required for solving the corresponding problem. Our capability to use global context for reducing ambiguity without the formalization is connected with peripheral consciousness. It takes into account the hidden guiding lines and some grammatical constructions, which are contained in the context. This mechanism must be exactly formulated for a machine [135].

The pronounciation and understanding of a natural language sentence assumes the informational relation between the sentence and the context. In order to force the machine to understand speech and accomplish the speech communication problems (such as translation, speech control, etc.) it is necessary to teach the machine to relate the words and phrases with real situations. There are two approaches to the decision of this problem. The first approach is based on the scrupulous description of the object domain, i.e. on the universal knowledge presentation method. For example, *The chair is furniture, it serves for sitting*, etc. Such description is inconvenient and it is difficult to make a conclusion based on such information. The second approach is derived directly from the speech acts theory. This approach is based on the situational model of the activity and it is the subject of our studies (Section 4).

2.7. Main challenges of speech processing

Based on the analysis of most research in speech man-computer interaction domain we can select the following key challenges:

- Creation of robust speech understanding. Here the human perception mechanism may serve as a reference point, which ignores many types of speech variability and accomplishes "the correct meaning recognition from incorrect words".
- Creation of models for integration of all kinds of information during the speech understanding process.
- Rational completeness degree of input information. Traditional methods of semantic analysis are usually based on the following idea: "Understanding is the capability to respond to all questions connected with input utterance". This principle is usually reduced to the necessity of scrupulous filling all semantic frame slots that leads to the complexity of the dialogue management module, reentry of words or utterances and groundless delay of dialogue progress. A human intuitively feels the situation and understands different reductions, idioms, and professional jargon. In other words there is some reasonable completeness level for dialogue phrases sufficient for their reliable understanding without any reentries. So statistical data, which sufficiently completely reflects human speech behavior may be useful for the decision of this problem.

- Improvement of word recognition modules: speech recognition accuracy, robustness to phonetic and grammatical deviations, rapid adaptation to the speaker's voice, acoustic environment and specific application domain.

3. The approach of Speech Informatics Group to the speech dialogue problem

During the elaboration of the speech understanding model the group was aimed at achieving maximal accuracy and robustness due to using the hypothesis of integral data processing. This hypothesis was taken from the text processing area and was essentially developed and modified for the speech understanding task. In this section the mathematical base of the model is presented. The base premises and analogies were taken from human speech perception, the speech acts theory and the communication theory. The integral model contains three levels of processing: acoustical, associative, pragmatic. In this section high-level processing is considered in detail.

3.1. Conceptual premises. Human speech perception

The group uses the following principles in speech understanding research:

- Human speech perception can serve as an ideal prototype for the natural spoken language (NSL) modeling. The main sources of knowledge about the properties of this system are: psycho-acoustics of speech [153], areas of psychology concerning speech errors and defects [161], areas of neuro-physiology, dedicated to speech and language processing [66].
- The main listener's goal during the speech communication process is to guess the speaker's intentions (but not to calculate the truth or falseness) of utterances. The minimal unit of the human speech communication is not a sentence or any other speech expression, but it is an action, which is the meaning of this sentence [5].
- The process of speech perception contains the procedures of forming the sounds, words and meaning hypotheses, their quantitative estimation of conformity with the concrete a priory data (acoustic-phonetic, lexical, syntactical, semantic and pragmatic data) and at last, producing semantic interpretation based on some integral estimation.
- An attempt of semantic interpretation must be present always, in spite of the doubts arising on certain levels of analysis.
- The information of syntactical and semantic types should be processed jointly as it corresponds to human nature [101,103]. Besides, according to the most preferable hypothesis all this processing is realized in the human's sub-consciousness by the associative mechanism [19] so it is expedient to use this hypothesis in the model.

- The current real situation should be taken into account for the improvement of understanding of a speech message. More adequate NSL model should include a model of the object domain.
- The role of language rules within the model should be reduced during the transition from 'right' speech to real spontaneous speech.

3.2. Associative analysis

Associative analysis of words chains (phrases) is based on the following assumptions:

- Semantic and syntactic relations are realized in human sub-consciousness (and in the brain) by the associations mechanism in the same process [70,81].
- The associative connection between two words can be evaluated by using the bigram statistics [22,59] or expert estimations [24,39].
- Different words chains can be estimated according to their degree of relationship based on inter-word associations [52].

Let the vocabulary be $W = \{w_1, w_2, ..., w_g, ..., w_N\}$. The matrix with dimension N*N, $a \geq 0$ $A_{[N,N]} = \|a_{gh}\|$ contains coefficients for every ordered words pairs (w_g, w_h). Experts build the matrix using the discrete scale (for instance, 4-scores). For any word chain (phrase) with the length $L(f_n = w_{n_1}, w_{n_2}, ..., w_{n_i}, ..., w_{n_L})$ the subset A^*, which contains coefficients for every ordered word pairs of this phrase, is extracted from matrix A:

$$A^* = \{a_{n_1,n_2}, a_{n_1,n_3}, ..., a_{n_2,n_3}, a_{n_2,n_4}, ..., a_{n_{L-1},n_L}\},$$

$|A^*| = C_L^2.$, where C_L^2 is the amount of obtained combinations.

As a result the associative measure of the phrase f_n is the sum normalized by the length L:

$$E_{ass}(n) = \frac{1}{C_L^2} \sum_{n_k=1}^{L-1} \sum_{n_s=n_k+1}^{L} a_{n_k,n_s}, \quad k < s.$$

3.3. Pragmatic estimation in speech processing

It is known that complete speech understanding is possible only by consideration of a sufficiently wide situational context in connection with the sender's/receiver's life experience and their professional skills. We use the following suppositions: (1) each speech act is accomplished within a respective situational context, (2) a purposeful human activity in a certain applied area can be presented by the ordered structure of situations. It meets the human nature, and situational analysis is simpler than other kinds of pragmatic processing. In particular, being mentally in a concrete situation, people use a very restricted set of sentences in comparison with the full language model. It raises the robustness and efficiency of speech

processing. Moreover, from the point of view of high-level linguistic processing, many topics are simplified here, such as the influence of homonymy, antonymy, polysemy, ellipsis, anaphora, discourse, etc. For example, speaking about homonyms, it is unlikely to meet equally sounding words «stolb» (a telegraph-pole) and «stolp» (a pillar of society) in the same situation. The situational context accompanies all the dialogues: we understand each other well when we are both mentally «included» into the corresponding situation.

The purpose of pragmatic processing in our work is to estimate the degree of an input phrase correspondence to the current situation by the comparison of input hypotheses with possible canonical phrases. The fragment of the situational database presented in Table 2 contains the information on possible transitions and possible phrases producing these transitions.

Table 2. Fragment of situational database

Current situation	Following situation 1.	Phrase 1.1. Phrase 1.2.	Weights of words Weights of words . . . . .
	Following situation 2.	Phrase 2.1. Phrase 2.2.	. . . . .
	. . . . .	. . . . .	. . . . .

To avoid the total enumeration of all the possible phrases in a language model by their comparison with the input utterance, we use the quantitative comparison like it is made on the acoustic-lexical level. In this case the right solution can be found not only by exact coincidence, but also with all the deviations, which do not exceed a certain value. Both phrases are presented as sets of words. The words of the canonical phrase have some weights. A semantic difference is calculated using the regular machinery of the sets theory taking into account lengths l_1 and l_2 of phrases and weights of words:

$$E(Phr_1, Phr_2) = F(l_1, l_2, w_1, w_2, \ldots, w_s)$$

Two streams of hypotheses are processed on the pragmatic level (Figure 17): on the one hand, the hypotheses of the input message, selected according to the acoustic-lexical information, on the other hand, hypotheses of possible utterance meaning, generated according regularities of the applied area. Each acceptable type of meaning points out to a respective set of equivalent phrases [54,135]. Of course, such set can be presented as a common list or by some generative model.

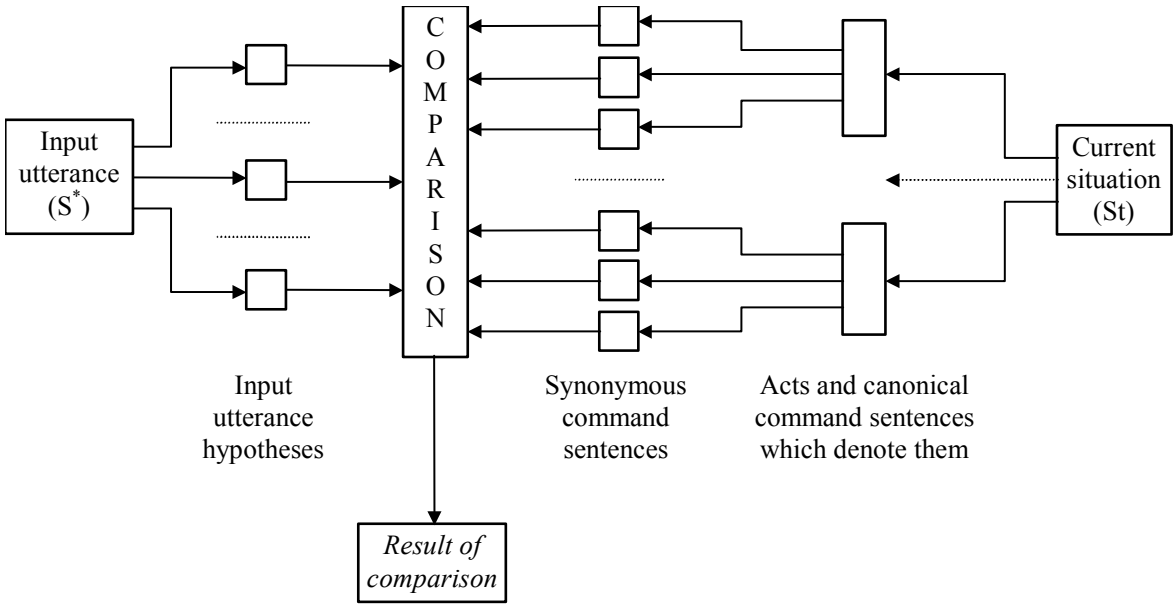


Figure 17. Pragmatic estimation of hypotheses of an input utterance

The estimation of correspondence between an input hypothesis and a concrete act in the framework of a current situation is obtained by quantitative comparison of the input hypothesis with the subset of equivalent phrases for the current situation.

$$St = \{St_1, St_2, ..., St_b, ..., St_B\},$$

where B is a number of situations, in which it is possible to transit from the current situation. Canonical command K_b exists for each situation $\mathrm{St_b}$:

$$St_b \rightarrow K_b$$

A certain subset of synonymous sentences exists for each K_b:

$$K_b \rightarrow \{K_{b1}, K_{b2}, ..., K_{bj}, ..., K_{bJ}\}$$

The semantic distance between input sentence hypothesis $\mathrm{f_n = H}$ and a canonical sentence $K_b = K$ is calculated using the following assumptions:

- H and K sentences are presented as subsets of words, because in natural speech the word order is often disordered.
- Each word $\mathrm{w_i}$ from the command has semantic weight $\mathrm{v_i}$ obtained by experts.
- The sum of all the words weights for each phrase is constant:

$$\sum_{i=1}^{L} v_i = const.$$

Then it is possible to present a command sentence K as a non-regulated set of pairs <word, weight>:

$$K \rightarrow \{< w_1, v_1 >, < w_2, v_2 >, ..., < w_i, v_i >, ..., < w_l, v_l >\}, L = |K|$$

It is practically impossible to evaluate weights for each words of an unexpected input utterance. Therefore the hypothesis about an input utterance can be presented as:

$$H = \{w_1, w_2, ..., w_i, ..., w_M\}, \quad M = |H|.$$

It is assumed that semantic discrepancy between the sets K and H depends on concrete lexical contents of K and H, as well as on L, M, and V_i. This discrepancy is estimated by using differences between sets $A_1 = H \setminus K$ and $A_2 = K \setminus H$, and the intersection $A_3 = K \cap H$ (Figure 18).

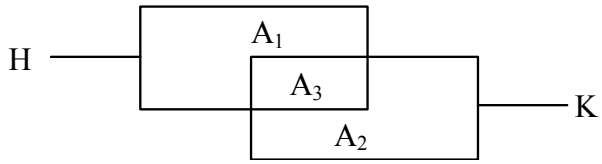


Figure 18. Semantic discrepancy between an input utterance hypothesis and a canonical sentence

Thus, for pragmatic difference the measure between the input phrase hypothesis and a canonical phrase the following formula obtained by empirical way is used:

$$D(H,K) = \frac{p_1|A_1| + p_2|A_2|}{L+M}\left(\sum_{i \in A_2} v_i + 1\right).$$

It can be seen that value D(H,K) increases when $|A_1|, |A_2|$ or the sum of words weights contained in subset A_2 increase. The multiplier $\frac{1}{L+M}$ is used to normalize this function by the summarized phrase length. The weight coefficients p_1 and p_2 are obtained by experts.

Pragmatic estimation is computed for each hypothesis of the input phrase. The hypothesis is compared with all the variants of canonical phrases suited to the current situation using the above formula. Then the optimization is accomplished i.e. minimal estimation is given to every hypothesis *n* and index *b* corresponding to a certain act.

$$E_{pr}(n,b,j) = \min_{n,b,j} D\{H_n, K_{bj}\}$$

3.4. The integral processing as the base of robust speech understanding

The definition of integral processing given by Schank and Selfridge [99,102] is based on the hypothesis that syntactic and semantic analysis (and also semantic and pragmatic) is performed at the same time and using the same mechanism. The requirement of mutual consideration of diverse kinds of knowledge is shown by the example of phrases processing (Table 3). It is assumed here that every kind of processing gives an estimate on 10 scores scale:

Table 3. Example of phrase processing by diverse kinds of knowledge

Phrase hypothesis N	Estimations		
	Acoustic-lexical	Semantic-syntactic	Pragmatic
1	10	0	0
2	7	8	8
3	9	6	5
4	0	0	10

Of course, a question arises which is the best hypothesis. It is obvious that in such case the knowledge integration allows to solve this problem by optimization methods.

As a result of the research the following definition was obtained: the integral processing is such one which allows to consider weighted results of partial processing for the possibility to get the optimal solution of speech understanding. This definition leads to the model structure shown in Figure 2 (right).

The mathematical presentation of integral processing is shown below.

$$\begin{cases} Knowledge \quad 1: \quad F_1(S) = 0; \\ Knowledge \quad 2: \quad F_2(S) = 0; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ Knowledge \quad n: \quad F_n(S) = 0. \end{cases}$$

Every kind of knowledge *1,2, . . . n* allows to obtain the estimations of the correspondence measure between an input utterance and a certain kind of knowledge. For example, at the word recognition level it is DP- or HMM estimations, at the semantic-syntactic level stochastic estimation can be used. It is obvious that zero estimations can be the ideal case. In general case during real speech processing the right parts of these equations will contain some residuals:

$$\begin{cases} \mathrm{F_1}(\mathrm{S}^*) = \mathrm{E_1}; \\ \mathrm{F_2}(\mathrm{S}^*) = \mathrm{E_2}; \\ \ldots\ldots\ldots\ldots \\ \mathrm{F_n}(\mathrm{S}^*) = \mathrm{E_n}. \end{cases}$$

Hence the vector $E=E_1, E_2, \ldots E_n$ may characterize the quality of an input hypothesis. Its length reflects the degree of deviation of the input hypothesis from a meaning hypothesis and can be used as the base for the optimal decision search according to the "minimal deviation" principle.

Thus the integral estimation is accomplished by linear combination of all estimations according to the following equation:

$$E_n = \left[\alpha_1 E_{ac}^2(n) + \alpha_2 E_{ass}^2(n) + \alpha_3 E_{pr}^2(n, j_n)\right]^{\frac{1}{2}},$$

where α_1, α_2, and α_3 are weight coefficients and j_n is the corresponding act number. E_n shows the integral deviation measure for every hypothesis. The hypothesis with minimal E_n is the final result of the integral understanding process.

Thus, the elaborated integral speech understanding model takes into account a number of important facilities of human speech perception as well as it is well-coordinated with the fundamentals of the information theory. Besides it contains acoustical-lexical, semantic-syntactical and pragmatic processing levels and forms the united integral estimation.

4. Theoretical investigations conducted during the project

This section is devoted to theoretical research conducted in the framework of the project. In order to create a robust and competitive model we researched the various levels of speech signal processing and carried out a wide spectrum of works.

In order to extract the speech signal from noise environments the method based on spectral entropy analysis was investigated and developed. The research of the parametrical representation of the speech signal was conducted during the project. Two methods (sign autocorrelation function [56] and spectral-difference features [139]) robust to variations of the signal level have been developed. Besides, the second method has shown the robustness to accidental nonlinear spectrum deformations.

During the project we have proposed and elaborated the sliding analysis method of the continuous speech recognition, which is robust to grammatical deviations in a pronounced phrase and has acceptable complexity. The method is based on multi alternative choice of the word hypotheses, posterior phrase hypothesation from these words, and their estimation taking into account acoustic probability, time probability and duration of a hypothetical phrase.

Special attention was paid to robustification of a speech dialogue model. In this section it is shown that the robustness of human speech perception is much better than that of the machine. Some examples of intuitive comparison of some methods for speech understanding are adduced. An attempt is made to offer the definition of robustness degree within the framework of the integral model, which is elaborated in the group.

Accuracy and flexibility of a system depend on its capability to adapt to various aspects of the application. Moreover during the debugging and exploitation the system can obtain the new data and so the database adjustment is necessary. For this aim the integral database adjustment was developed. The integral approach takes into account acoustic aspect, language aspect, subject area and also the optimization of model parameters.

As a result of the research the following modules have been created (1) the integral adaptation module providing the inter-coordinated adjustment of all databases of the understanding model; (2) the module of continuous speech recognition robust to grammatical deviations based on the sliding analysis of a speech signal and a posteriori phrase hypothezation. The developed modules have been included into the integral understanding model. As a result the united software complex has been created. It provides robust understanding of continuous speech as well as portability with respect to new applied tasks due to integral adaptation.

In the conclusion of this section the problem of extra linguistic information and its paramount importance in speech understanding process is discussed. At present approaches to formalization and usage of situational information are poorly developed. Therefore the problem of using situational information leads to the problem of creating situational databases, which contain constraints of the applied area. In our works the model of applied area for the restricted tasks connected with the control of technical objects (a car, a plane, a robot, etc.) is proposed. But there are no techniques of the creation of situational databases for other intellectual applications of speech technologies.

4.1. The development of the method for robust speech endpoint detection based on the entropy of the signal spectrum

At present there are many methods for speech endpoint detection based on the calculation of short-time signal energy, spectral energy, the number of zero-crossings of the speech signal, adaptive threshold values and information about the duration of speech fragments. However, all these algorithms become less reliable in conditions of non-stationary noise as well as during the appearance of diverse sound artifacts (aspiration, lip smack, etc).

Therefore the effective method for speech endpoint detection, which allows selecting the speech in conditions of non-stationary noise as well as during the appearance of sound artifacts, has been proposed.

During the development several main requirements to the method were made:

- minimization of the probability of false alarm caused by high level noise only;
- high probability of the correct detection of speech even in noisy conditions;
- high processing speed for the exclusion of delays while switching-on and switching-off speech recognizer.

4.1.1. Mathematical foundations of the method

The developed method is based on the calculation of entropy (as the measure of uncertainty or disorder in a given distribution [2]) of the signal spectrum. The distinction between entropy for speech segments and entropy for background noise is used for speech endpoint detection. Such criterion is less sensitive to the variations of the signal amplitude. For the first time the spectrum entropy was used for this task several years ago [2,104,116]. Our method is the development of those ideas and includes some new levels into the analysis of the speech signal. Figure 19 shows the block diagram of the developed algorithm for speech endpoint detection.

The algorithm functions as follows. The signal incoming through a microphone is digitized with sampling frequency 16 KHz and divided into short segments with duration 16 ms in each (256 samples in each segment). The neighboring segments have the overlap in 70 samples (about 27%).

Additionally at this stage diverse methods for the signal refinement from noise can be used (for instance, adaptive Kalman filtering or methods of spectral subtraction) [27].

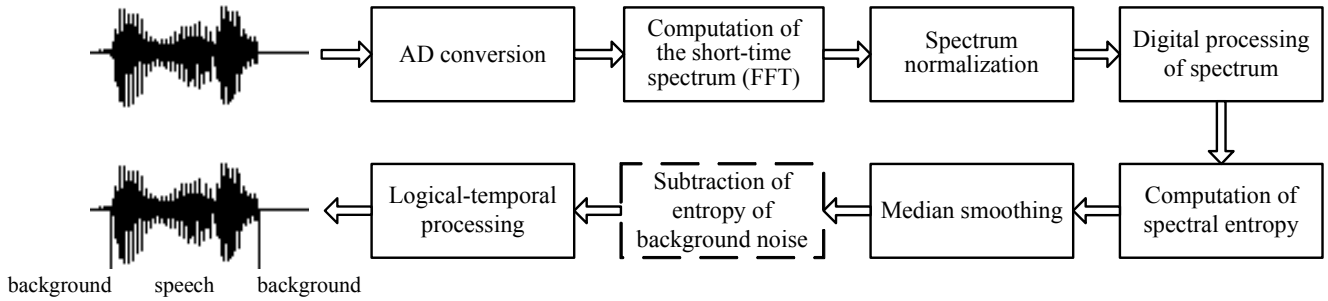


Figure 19. The algorithm for speech endpoint detection based on the entropy of the signal spectrum

Then the short-time signal spectrum is computed using the Fast Fourier Transform algorithm (FFT), and the normalization of the calculated spectrum over all frequency components is fulfilled:

$$p_i = \frac{s(f_i)}{\sum_{k=1}^{N} s(f_k)}, \qquad i = 1 \ldots N$$

where $s(f_i)$ is the spectral energy for the frequency component f_i, p_i is the corresponding density of probability, and N is the total number of frequency components in FFT. The calculated function is the probability density function. The number of frequency components can vary from several tens to several hundreds. Here a trade-off between the required sensitivity and the computational load must be found. In our model 256 frequency components are used.

To reject some kinds of noise at this stage of processing two restrictions are used:

- The frequency band is restricted from 200 Hz till 8000 Hz, i.e.,

  $s(f_i) = 0$, if $f_i < 200$ Гц.

  This frequency band covers most of the frequency components, which are used in human's speech. Such restrictions allow to exclude the influence both low-frequency noises and high-frequency noises (for instance, internal noises of a soundcard or a microphone).

- Acceptable values of probability density are restricted both from the above and from the below which allows to exclude the noises concentrated in the narrow band as well as the noises, which have approximately equal distribution of frequency components in the spectrum (for instance, white noise).

  $p_i = 0$, if $p_i < \delta_2$ or $p_i > \delta_1$,

where δ_1 and δ_2 are upper and lower boundaries of probability density correspondingly. In our model $\delta_1 = 0.3$ and $\delta_2 = 0.01$.

At the next stage of processing the computation of spectral entropy from the normalized spectrum is fulfilled according to the formula [104]:

$$H = -\sum_{k=1}^{N} p_k \log p_k$$

The median smoothing of sequence ξ of calculated values of spectral entropy is used at the next stage of the analysis. In contrast to many other smoothing methods (for instance, the method of sliding middles) this method is significantly more robust to the presence of outliers and other occasional data distortions. The calculation of the sliding median is the base of this smoothing method. To find the value of the sliding median at a time point *t* the median of the sequence in time frame (window) *[t-q, t+q]* is calculated. The median of the sequence in time frame is defined as the central member of the ordered (in increasing order) set of values, which are included in this time frame. During our experiments the best results have been obtained by the method of median smoothing in the window with size 5.

For some tasks, where the kind of noise and its spectrum are slowly changed during a long time, the additional computation of spectral entropy for background noise and the subtraction of this value from obtained entropy for an analyzable signal can be useful.

The logical-temporal processing, which takes into account the practically acceptable durations of speech and non-speech fragments, is used at the last stage. Firstly, the adaptive threshold, which is used for the detection of endpoints (beginning and ending) of the hypothesis of the speech fragment, is calculated:

$$\gamma = \left(\frac{\max(\xi) - \min(\xi)}{2} + \min(\xi) \right) * \mu ,$$

where μ is the coefficient, which is chosen by an experimental way. In our model this coefficient takes values from 0,8 till 1,1 depending on the level of noise of the signal. The threshold γ has a minimal value, which allows to detect the speech segment exactly. In our system the value of the threshold is 1,6 (at values of spectral entropy from 0 till 3). Based on this threshold the acoustic segments of the analyzable signal, which belong to a human's speech, are selected. Then the logical-temporal processing of selected signal fragments is fulfilled. This processing is required since in many cases the non-speech signal fragments are recognized as speech owing to appearance of diverse sound artifacts. And vice-versa, some fragments, which contain speech, are rejected. For logical-temporal processing two criteria are used (Figure 20):

- The duration of the selected fragment with speech – S (regions a_ib_i and a_jb_j)
- The duration of the interval between two neighboring selected fragments – N (b_ia_j)

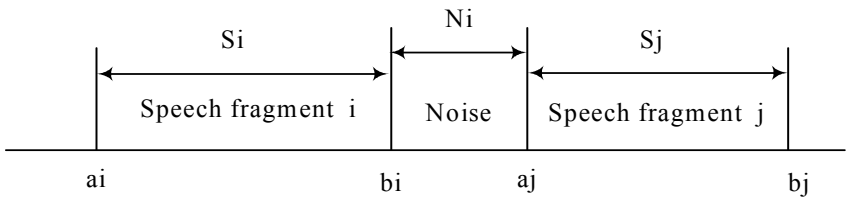


Figure 20. The essence of logical-temporal processing

Taking into account that a human cannot produce very short speech fragments as well as that there are always some pauses in speech (for instance, before explosive consonants), we have experimentally determined optimal values for S and N (15 and 20 segments correspondingly). Thus, if some part of a signal meets the above-mentioned requirements then selected fragments inside this part are fused into one speech fragment (region $a_i b_j$). And this fragment of the signal is the result of work of this algorithm.

4.1.2. Experimental results

Some experimental results for the developed method are presented below. The method has been tested by separately pronounced words and continuously pronounced phrases from background noise, in which diverse kinds of noise were artificially included (by means of the program CoolEdit Pro 2000). The following kinds of noises were used in our experiments:

1. Noise with narrow band (2700-3300 Hz). This noise can be approximately considered as monotonous signal with frequency 3000 Hz.
2. White noise. This noise has the spectrum with approximately constant spectral density in the frequency band 0-8000Hz.
3. Brown noise. The spectral density decreases by 6 db with each subsequent octave (i.e. spectral density is inversely as the square of frequency).
4. Pink noise. The spectrum of such noise has spectral density, which decreases by 3 db with each subsequent octave (i.e. spectral density is inversely of frequency).
5. Amplified acoustic background noise recorded in a room where all experiments were made.

Figures 21 and 22 show the example of the selection of the word “attention” from the signal, which contains all above-mentioned kinds of noises with amplitudes, which are higher or approximately equal to the amplitudes of pronounced speech. It is clear that the speech fragment in this signal has been selected absolutely correctly. It can be noted that the algorithm has coped with all kinds of noises (i.e. did not recognize these noises as speech).

Figures 23 and 24 show the example of the selection of the word “attention” from the signal, which has been obtained by mixing the original sound signal with pseudo-random white noise with high amplitude. In this experiment the signal-to-noise ratio (SNR) was about 3 db. It is clear that the algorithm has accomplished this test correctly.

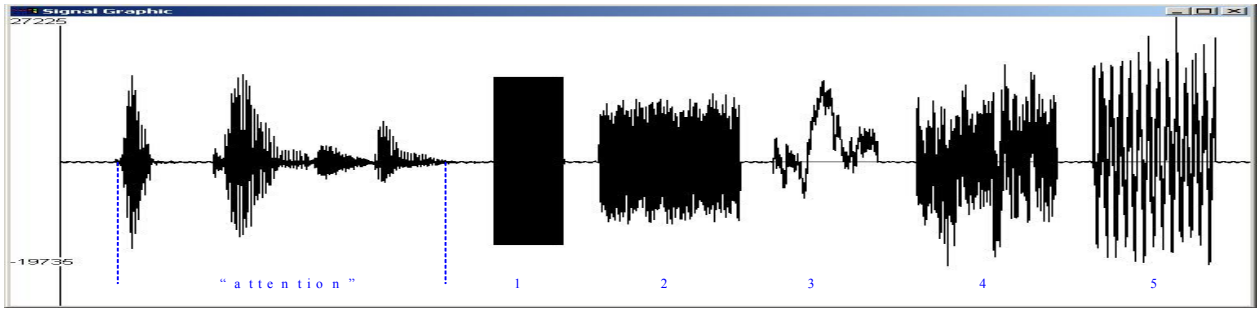


Figure 21. Test signal

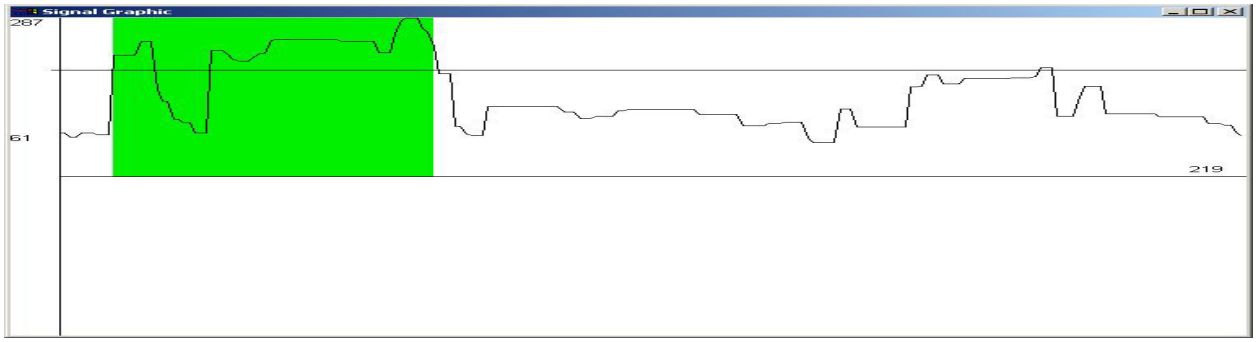


Figure 22. Example of speech endpoint detection

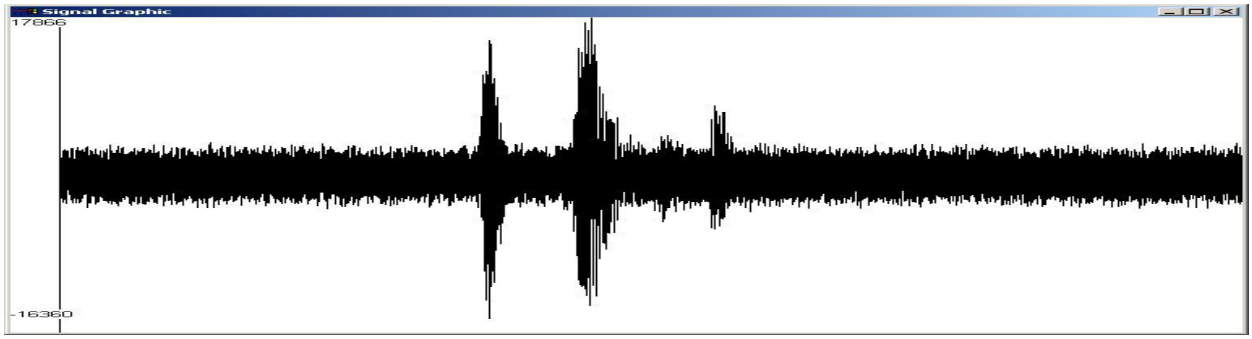


Figure 23. Test signal with additive noise

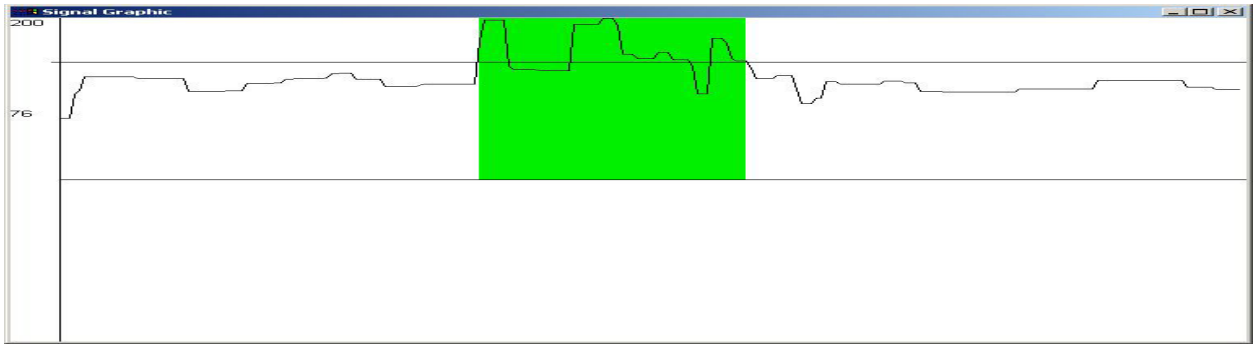


Figure 24. Example of endpoint detection of speech with additive noise

Besides, the measure of efficiency of the developed algorithm is the error rate of speech endpoint detection [147]. This rate consists of:

- The rate of false alarm R_{fa}, i.e. the detection of the speech message in the point of time, when there is no speech activity.
- The rate of truncation of the speech fragment R_{st}, i.e. in spite of the presence of speech at a time point the algorithm cannot detect any speech fragment.

Thus the error rate of speech endpoint detection is $R_{sed.err} = R_{fa} + R_{st}$

Table 4 presents the results of our experiments according to automatic detection of separately pronounced words (150 English words) in the signal, in which diverse kinds of noises have been inserted.

This table shows high effectiveness and robustness of the algorithm in conditions of non-stationary noises (the worst case is the presence of pink noise, the form of which looks like a real speech signal).

Table 4. The detection of speech at the presence of diverse kinds of noises in the signal

Kinds of noise		**Rate of false alarm, %**	**Rate of truncation of speech, %**	**Error rate of speech detection, %**
1	Narrowband noise	0	2	2
2	White noise	1	2	3
3	Brown noise	3	2	5
4	Pink noise	15	3	18
5	Amplified background noise	2	2	4

The error rate concerning the truncation of the speech fragment remains almost constant for all experiments. It can be explained by periodical truncation of the pause before explosive consonants (“t”, “d”, “p”) in the beginning of a word. This fragment is often recognized as silence (or background noise). However at present there are not any efficient methods, which could solve this problem.

The main difficulties of the task are caused by unsteadiness of pronunciation, the presence of specific pauses inside words, influence of non-stationary noise.

The experimental results of the usage of the developed method have shown that speech fragments are successfully selected in sound signals, which have diverse kinds of intense noises and sound artifacts. Moreover, the developed method has sufficiently high speed of processing and can be used in real-time (on-line) speech recognition systems.

4.2. Development of the features robust to variations of signal scale and accidental nonlinear spectrum deformations

The important problem, which must be decided by developer first of the all, is elaboration of the method for **parametric signal representation**, which could exactly differ the phones and words of speech and at the same time be invariant to pronunciation peculiarities of the concrete speaker, change of the acoustic environment, microphone, etc [90,129,140].

There are many methods for the parametric signal representation based on various spectral transformations, regression analysis, autocorrelation analysis, etc. But it is sufficiently difficult to choose the best method among them. One of the essential disadvantages of the known methods is instability to signal amplification that significantly decreases the quality of speech recognition. Research of the parametrical representation of the speech signal was conducted during all time of the project. As a result two methods (sign autocorrelation function [56] and spectral-difference features [139]) robust to variations of the signal level have been developed. Besides the second method has shown high robustness to accidental nonlinear spectrum deformations. Below this method is described in detail.

4.2.1. Problem definition

At present time many feature systems are used, which are equivalent at recognition rate. The most popular of them are spectral and LPC features.

Many word recognition errors arise because of the inaccurate position of the microphone or the voice volume change. The other cause of errors is the nonlinear deformations of signal spectrum form.

One of the most important tasks in creation of robust recognition/understanding systems is a choice of the method for parametric signal representation, which is adequate enough to its content and at the same time invariant (as far as possible) to the deviations of signal scale and to accidental spectrum deformations.

During the history of creation of speech recognition systems several variants of feature systems were proposed, which meet the mentioned requirement. One of the directions in the search of such features is the one based on the comparison of spectral bands energy. The founder of this direction is L.L. Mjasnikov. He used the feature system robust to the variations of the signal scale, in which the ternary coding was applied for the phoneme recognition [150]. In his electronic device the speech signal was yielded to the band-pass filter bank. Then the energy of spectral bands is compared in pairs and the features with values +, -, 0 were formed. Depending on the combination of obtained features one or other relay adjusted to the concrete phoneme worked. These features allowed to recognize “pulled” sounds enough surely. The features similar to the described above by formation principle were afterwards used for the isolated word recognition [132]. In this recognizer the binary features were used, which were formed by means of sign comparison of the signal energy in some spectral filters. The sign “+” is generated for the areas of increasing and “-” – for the areas of decreasing. The disadvantage

of this method is the invisibility of areas without the tilt (for instance, the zero signal) and the areas with the negative tilt that leads to errors. The ternary features were used for isolated word recognition too and they showed better results than binary ones [143]. On the one hand all these methods are robust to the influence of linear and nonlinear deformations of the signal spectrum, but on the other hand they are too rough, because they do not take into account the steepness of the spectral function in different areas.

During the Project we have proposed the new class of features named the spectral-difference features (SD features), which are invariant to the signal scale and take into account variations of the tilt steepness of the spectral function.

4.2.2. Representation of features set by means of piecewise-linear approximation

Let a discrete signal spectrum is $A = a_1, a_2, a_3 ... a_n$.

To escape the linear deformations of the spectrum (change of signal level) it is enough to use the relative coefficients $K_1, K_2, ..., K_{n-1}$: $A \to a_1 = K_1 a_2,\ a_2 = K_2 a_3, ...,\ a_{n-1} = K_{n-1} a_n$.

However, such description could not decide the problem of nonlinear deformations, which are permanently present in speech signal at pronunciation of the same sounds by the same speaker. For taking into account such accidental nonlinear deformations we use the permissible deformation zone $\pm \delta_i$ for each approximation area (Figure 25).

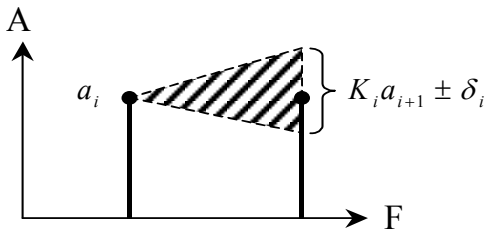


Figure 25. Permissible deformation zone

The equations set: $a_i = K_i a_{i+1} \pm \delta_i$ is transformed into the set of inequalities: $\begin{cases} a_i \geq K_i a_{i+1} - \delta_i & (1) \\ a_i \leq K_i a_{i+1} + \delta_i & (2) \end{cases}$, which define one SD feature. If both conditions hold true then the feature value is “equal”. If only condition (1) holds true then the feature value is “more”, if only condition (2) is true then the feature value is “less”. A simultaneous non-fulfillment of both inequalities is impossible. Thus each spectrum area is coded by ternary code, for instance $\{0,1,2\}$.

Then the main problem is the choice of the optimal set of features. The relative coefficients and dimensions of permissible deformation zone must be optimized too. The optimization procedure is described in detail below.

4.2.3. Forming the spectral-difference features

The base for SD features computation is the procedure of obtaining the cepstral features, which is the most popular kind of spectral features.

The sequence of operations, used for transformation of speech signal into spectral-difference vectors is presented in Figure 26.

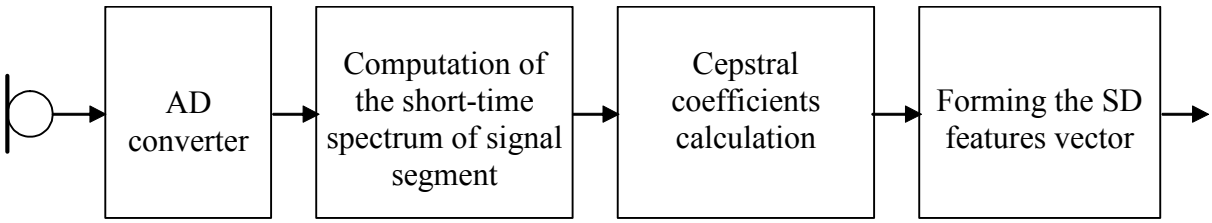


Figure 26. Procedure of obtaining the spectral-difference features vector

We digitize the speech signal, incoming through a microphone with sampling frequency 16 KHz and divide it into short segments of 186 samples in each (about 11 ms). During the analysis the segment-rate is set to 86 segments per second. Selected segments come into the block of computation of the short-time signal spectrum. In this block the segment is multiplied by Hamming window [106]. It is done to escape the effect of significant "leakage" of energy, which can lead to masking the weak signal spectrum components, which may be observed when using the ordinary rectangular window. Then the short-time signal spectrum is computed using the Fast Fourier Transform algorithm (FFT). For the FFT 256 samples of signal are used, which are obtained by addition of 70 samples of the previous segment to 186 samples of the current segment. Obtained short-time spectrum, composed of 256 spectral samples, comes at the block of cepstral coefficients calculation. In this block a set of spectral coefficients, which characterize the signal power at the outputs of band-pass filter bank, is calculated. We use the bank of 8 overlapping triangular filters (Figure 27), the frequency values of which were chosen by experimental way (Table 5). These filters cover the frequency band from 200 Hz till 8000 Hz.

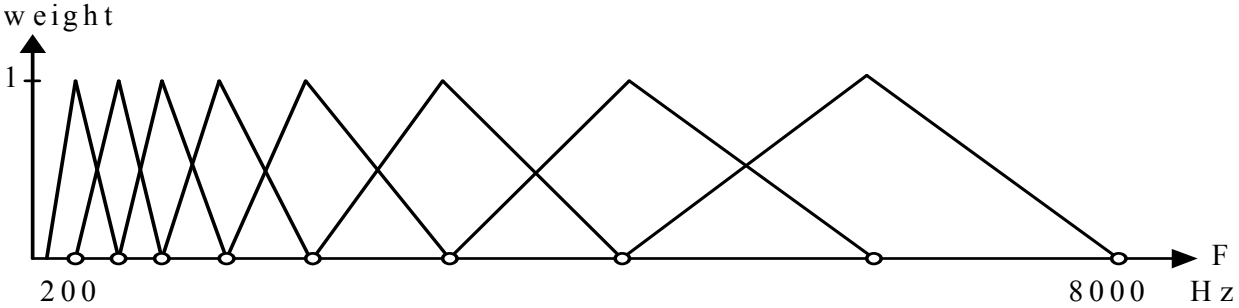


Figure 27. Bank of triangular filters

Table 5. Parameters of triangular filters

Filter number	Minimum value of frequency, Hz	Maximum value of frequency, Hz	Frequency value of the vertex of triangle, Hz
1	200	500	300
2	300	750	500
3	500	1050	750
4	750	1500	1050
5	1050	2300	1500
6	1500	3400	2300
7	2300	6000	3400
8	3400	8000	6000

At the output of the block the spectral coefficients are transformed into cepstral coefficients [91]. Then inside the block of forming the SD feature vectors the sign comparison of diverse pairs of cepstral coefficients is made. The comparison is performed by the method described above. Each feature is coded by two bits: 00 is increase area; 10 is decrease area; 01 is area without tilt or with minor tilt of spectral function (falling into the permissible deformation zone); the state "11" is not used.

4.2.4. Optimization of spectral-difference features system

The optimization of the feature system is required for the choice of the subset of the most informative features, which provide the required recognition rate. It can be performed by the method of sequential exclusion of components from the full feature set. The program for automatic features optimization has been developed. The order of optimization process is presented in Figure 28.

At the first phase the wittingly surplus number of spectral band pairs is chosen and for each pair the subset of inequalities, taking into account different variants of linear and nonlinear deformations, is defined. Thus each pair of inequalities allows forming a corresponding feature of the initial system.

In our case the initial feature system is obtained by complete enumeration of all possible pairs of cepstral coefficients taking into account relative weight coefficients: 1:1, 1:2, 2:1, 1:4, 4:1. Using n spectral bands $n(n-1)\backslash 2$ features can be calculated for each weight relation. Thus for 8 mentioned above spectral bands the initial feature system contains 140 ternary features. All speech signals, contained in the template set and in the test set, are coded by vectors of this feature system.

At the second phase the vectors come into the block of exclusion of the least informative features. At this phase the recognition rates of feature systems, obtained by sequential exclusion with the following restoration the one feature from the initial system, are compared. The recognition rate can be evaluated by one of two possible criteria: the number of recognition

errors or reserve coefficient in case of recognition error absence. As the optimization was performed on the small test set, the first criterion turned out ineffective and the following optimization was conducted by the second criterion. The reserve coefficient is defined as follows:

$$Cr = \sum_{i=1}^{N} \frac{\min\left(d(W_i', W_j)\right)}{d(W_i', W_i)},$$

where N is vocabulary size;

W_i' is test realization of word i, $1 \le i \le N$;

W_j is template realization of word j, $1 \le j \le N$.

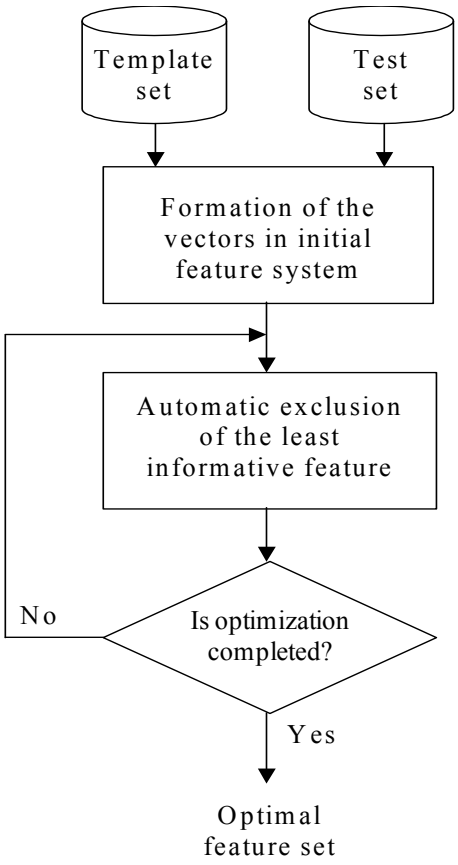


Figure 28. Optimization of feature system

The optimization procedure is performed as follows. Some feature is excluded and the resulting reserve coefficient is calculated. Then the feature is restored and the operation is repeated for all other features. As the result of the analysis of the obtained estimates the worst feature is excluded. The exclusion procedure is continued while the reserve coefficient is increased.

We have chosen 20 features from the initial system. Recognition rate was tested by the other set of 100 English words. The reserve coefficient has increased and the number of recognition errors has decreased. The recognition rate for the initial system is 95% and the recognition rate for the optimal system is 98%. It allows us to say, that optimal system of SD features is obtained.

4.2.5. Experimental results

We have evaluated the rate of recognition of isolated words using optimal set of SD features, test vocabulary with 100 English words and three diverse speakers. The system adjustment to a concrete speaker was performed by means of preliminary input of his templates. In this experiment the SD feature system was compared with two others: cepstral and autocorrelation features. The results are presented in Table 6.

Table 6. Comparison of recognition rate for some kinds of features

Speaker	Recognition rate, %		
	Spectral-difference features	Cepstral features	Autocorrelation features
1	98	96	92
2	97	96	90
3	97	95	89

It is obviously from the Table, that spectral-difference features provide the best recognition rate.

Evaluation of the robustness of features. We have tested the robustness of developed SD features to the variations of signal scale, which are the most frequent reason for errors. 100 English words were pronounced twice and saved by the speaker. The first pronounced set is the templates and the second is tests. The volume of test set of utterances was accepted as 100%. The recognition rate was tested by giving the original version of test utterances at the recognizer's input.

Then some experiments were made where the set of templates was invariable but the reproduction volume of the test set decreased consecutively. Then two tests were made using fourfold and eightfold reduction of the signal. The experiment has shown that the recognition process goes with high recognition rate in spite of the multiple reduction of the signal. Then the experiments with the consequent increase of reproduction volume of the test set were conducted. It must be said that when the signal level increases the process of signal cutoff at the maximum permissible level may be observed. In spite of the described effect the system was robust to the increase of the signal volume too. At that abrupt deterioration of recognition rate was observed only when the signal was multiplied more 3 times.

The results of accomplished research and their comparison with analogous experiments for cepstral and autocorrelation features are presented in Figure 29.

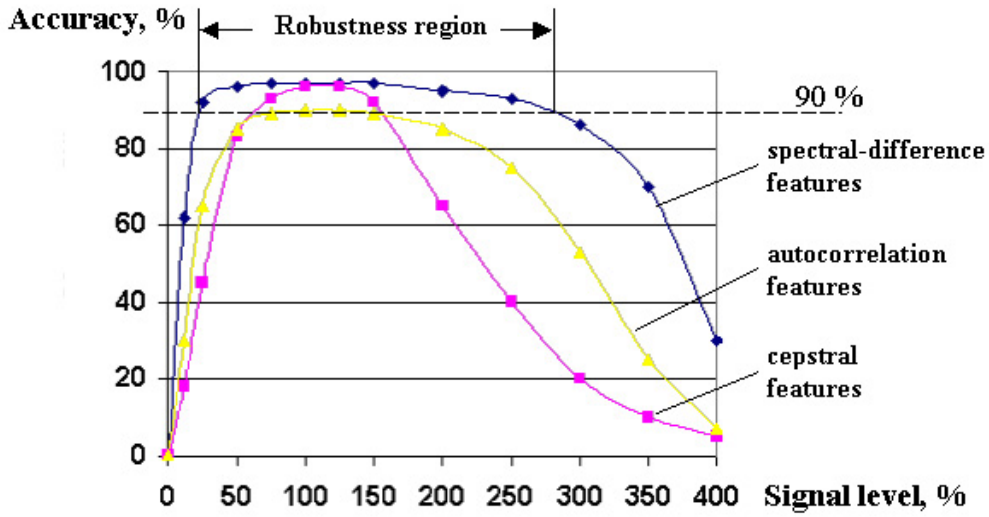


Figure 29. Comparison of robustness of spectral-based feature systems

It is obviously that developed spectral-difference features are more robust to variations of signal scale in comparison with other kinds of spectral-based features.

The stage of parametrical signal representation is main one in the preliminary speech processing and here it is important to extract all useful information from the speech signal. Therefore the research of optimal parametrical signal representation will be continued further.

4.3. Elaboration of the continuous speech recognition method robust to grammatical deviations

The development of continuous speech recognition models has been already going on for more than 20 years but acceptable decisions, which could be suitable for most important and perspective applications, are not found until now. The continuous speech does not have any separators, unlike the text. It is the main difficulty of the continuous speech recognition. So the problem of the word chain recognition is usually decided by the method of generation and verification of phrase hypotheses.

Practically all the approaches to continuous speech recognition are based on the principle of composite acoustic templates. The essence of the principle consists in the following: the hypothetical sequence of words is generated from a given vocabulary in one way or another. The rules of generation can vary from strict syntax to complete enumeration. Every hypothetical sequence is compared with the input signal by the dynamic programming (DP) method or by hidden markov modeling [46,64,79,98,131]. It is clear that the complete enumeration is only possible in very specific tasks, for instance in the recognition of the digit sequence [89]. But in cases when the vocabulary size is increased to hundreds and thousands of words, the speech recognition task becomes computationally unacceptable. When the vocabulary contains 100 words the number of four-word phrases hypotheses will be $(100)^4 = (10)^8$. So practically certain limitations are used. These limitations are based on strict syntax or stochastic n-gram model

[17,67,68]. But in this case the model is not capable of dealing with syntactically incorrect or stochastically unacceptable hypotheses. In order to make the recognition process robust to distorting factors [4,58,79], the generation model has to be less restricted and at the same time has to avoid complete enumerating.

One of the main aims of this project is the creation of methods for robust speech control. So the development of continuous speech recognition methods, which do not use any grammatical limitations, is required here.

Generally speaking the presented method is based on multi alternative choosing of word hypotheses within an utterance, generating phrases hypotheses from these words, and their estimation taking into account word acoustic likelihood and time intervals likelihood [57,96,94,156]. A more detailed description of this model is presented below.

4.3.1. The continuous speech recognition model based on sliding analysis

Continuous speech does not admit simple and unequivocal partition of its elements (phonemes, words, phrases), as these elements do not have obvious physical boundaries. Probably they are selected in the mind of a native speaker as a result of the complex multilevel process of speech recognition and understanding. The numerous attempts of preliminary automatic partition of speech have not yielded positive results.

We offer the approach, which eliminates the procedure of preliminary partition of speech and has a number of analogies with the processes taking place in neural structures. We take into account a hypothesis, according to which the brain reflects external influence as a space-time matrix of excitements [66]. The result of the acoustic recognition is the hypotheses subset of the word sequences with corresponding estimates of acoustic probability. These estimates are necessary for the following integral processing. Thus phrase hypotheses are built on the base of some preliminary analysis of the input signal, which significantly reduces their number compared to the generally accepted method of hypotheses generation. The size N of this hypotheses subset is chosen so as to minimize the loss of the correct hypothesis.

The structure of the acoustic level of the continuous speech recognition is presented in Figure 30. The input signal, which is extracted by the preliminary procedure of the proposition boundaries determination, enters the module of parametrical speech presentation. In this module the sequence of digital samples is divided into speech segments. A vector of parameters is calculated for every such segment. These vectors are transformed into the numbers of the vector templates by a vector quantization technique. Thus the base of the parametrical speech description is represented by the sequence of the numbers of template vectors.

The preliminary processed input signal enters to the set of the sliding DP-analyzers. Each of them is fitted on the concrete word template from vocabulary and searches the word hypotheses. Then united stream of all words hypotheses used for construction of phrase hypotheses. After that the estimation and ordering of phrase hypotheses are accomplished by acoustic and time parameters.

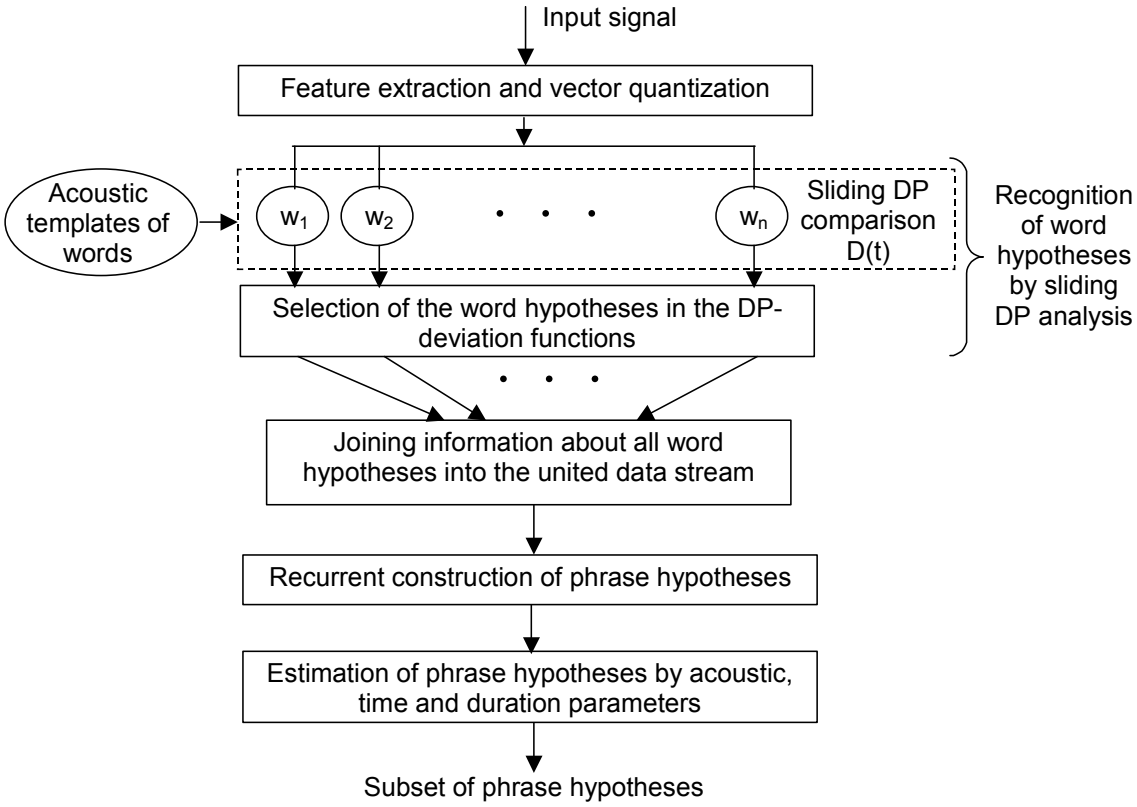


Figure 30. The structure of the acoustic level of the continuous speech recognition

Figure 31 shows the process of sliding DP-analysis. The word template slides along input signal with slide step and the DP-deviation between the template and signal part is calculated for every step. As a result we obtain the function of DP-deviations *D(t)*.

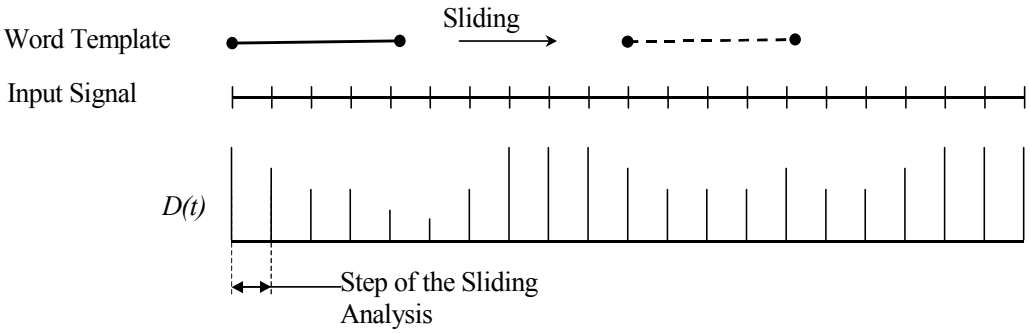


Figure 31. The process of comparison of the word template and the signal

The smaller DP-deviation between the word template and signal part the more probability of the word appearance. The Figure 32 shows the functions of the acoustic probability for N words. The local maximums of these functions are words hypotheses.

Then we construct all possible word chains hypotheses using the recurrent procedure of phrase hypotheses construction. The all words hypotheses discovered during sliding analysis enter to the input of the procedure. Recursion continues as long as the length *l* of hypothetical

phrase does not achieve the maximal acceptable length L. As a result the construction of the word chains hypotheses with length from 1 to L words is accomplished.

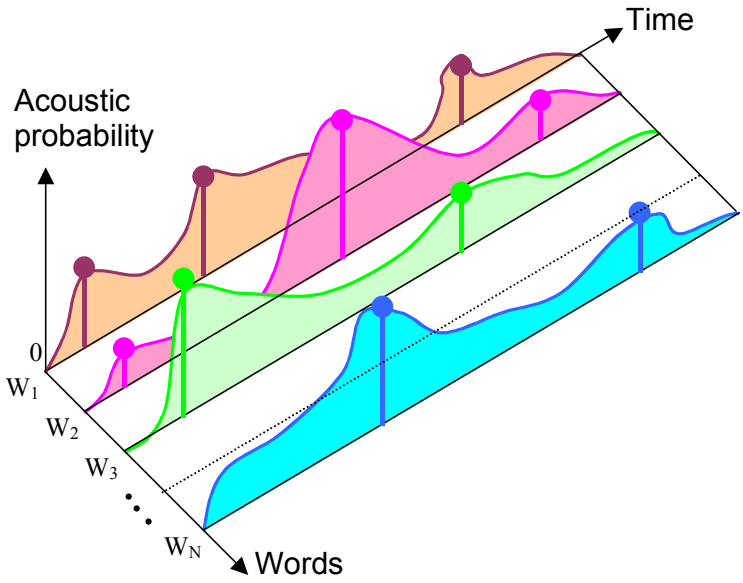


Figure 32. Acoustic probability of words appearances

Then every such chain (phrase hypothesis) is estimated taking into account its acoustical, time probabilities and total duration of word chain hypotheses. The acoustic probability is calculated during the sliding analysis as it was described above. The calculation of the time probability and total duration of word chain hypotheses is considered below.

The time probability is based on estimation of time intervals between the words hypotheses using the methodology of fuzzy sets [123]. According to the widespread practice for the B probability estimates we use the function of the belonging in the form of a trapezium. In our case the right side of the trapezium is not used, as unexpected pauses are possible between words in continuous speech. In Figure 33 the belonging function B is constructed for a word hypothesis, which is first and marked bold. Thus the belonging function B has the following three time intervals:

1) improbable intervals, where function B equals zero;
2) transitional intervals, in which the belonging function continuously increases from zero to one;
3) probable intervals, where function B equals one.

Four hypotheses of some words are shown in Figure 33. Let us consider three combinations of the hypotheses 1-2, 1-3, 1-4 and calculate the probability of their time compatibility. The second word is found in the first time interval of the belonging function of the first word. Therefore the combination probability of 1-2 words equals zero. The third word is found in the second interval of the belonging function and correspondingly the probability meaning of the 1-3 combination is found in the interval from zero and one (it approximately equals 0.6). Reasoning by analogy we obtain that the time probability of the 1-4 combination equals one.

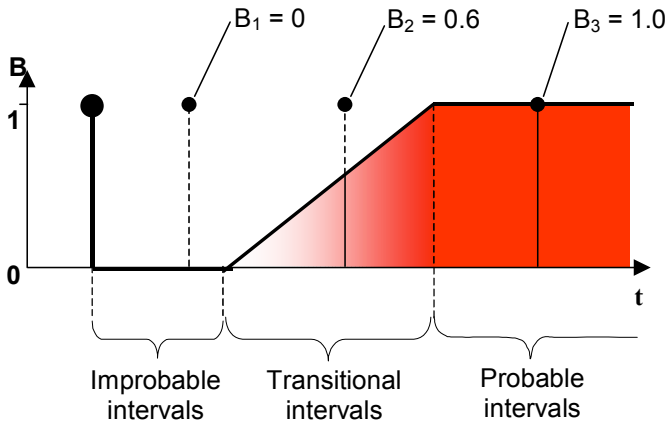


Figure 33. Temporal probability of the intervals between word hypotheses

In order to calculate the time probability of the words chain hypotheses we have to multiply time probabilities of the interval between word hypotheses contained in the chain.

Moreover during the research it was discovered that some false word hypotheses appear due to the mistaken matching of the template with “else's word” or its part. The information about the duration of the signal part, which is optimally matched with the word template during sliding DP-comparison, was not taken into account. At the same time importance of this information is obvious. We made an assumption that if the word hypotheses correspond to real words then the sum of corresponding durations of the parts will be close to the duration of the recorded speech signal. In case of mistaken similarity different deviations in the distribution of durations are probable, which will lead to the deviation of the mentioned sum from the signal duration.

The research based on such assumption was conducted and probability estimation of total duration of the hypothetical phrase was elaborated.

The probability of total duration of the hypothetical phrase is obtained in the following way: (1) the total duration of the phrase hypothesis is calculated by adding of the durations of the corresponding word hypotheses taking into account the overlaps and pauses between adjacent word hypotheses; (2) the correspondence measure $K(\tau_{ph})$ of the obtained duration and the duration of the input signal is determined. In the ideal case the words in the phrase must go word by word without overlaps and pauses. But overlaps and pauses between adjacent words are possible in natural speech. Therefore for calculation of $K(\tau_{ph})$ coefficient we use the “soft” evaluation; in which overlaps and pauses between adjacent words are estimated quantitatively.

Thus the estimate for the phrase hypothesis equals the multiplication of all the acoustic probability of word hypotheses, time probability of the interval between them and $K(\tau_{ph})$ coefficient. Then the united formula for estimation of H_{ph} phrase hypothesis consisting of l words presented in the following mode:

$$H_{ph} = K(\tau_{ph}) \cdot P_1 \cdot P_2 \ldots \cdot P_l \cdot B_2 \cdot B_3 \ldots \cdot B_l$$

As a result the set of word chain hypotheses and their probability estimates are obtained at the output of the acoustic recognition module of the continuous speech. A subset of the best hypotheses enters the following modules of high level processing. The interaction of the continuous speech recognition module with the integral speech understanding model is described in the following sections.

Thus the semantic-syntactic constraints are not used on the stage of phrase hypotheses construction, the model recognizes both correct phrases and phrases with some inaccuracies that saves the possibility of the further robust understanding.

4.3.2. Description of the model parameters

The developed algorithm of continuous speech recognition can be divided into the three stages: (1) sliding analysis of an input signal by a set of word analyzers (realized by software), which calculate the difference between the word template and various parts of the signal (phrase) by dynamic programming (DP); (2) selection of word hypotheses on the basis of local minimums of the DP-function and determination of their acoustic and time data (i.e. corresponding DP-estimates and appearance moments of local minimums); (3) construction of phrase hypotheses, their estimation on the basis of acoustic estimates of word hypotheses and time intervals between the moments of their appearances with the help of fuzzy sets theory.

It is obviously that the model is sufficiently complex and the analytical way of parameters selection is practically excluded. List of the main parameters, which are to be optimized, is presented below.

Every word from a vocabulary has its own analyzer, which contains the acoustic template and works in the following way. The signal selected by the algorithm of phrase boundary detection enters into the analyzer input. The process of comparison between the word template and the signal part is shown in Figure 31. The word template is shifted relatively of the input signal with AS step (a step of the sliding analysis). The DP-estimate is calculated at every step. The sequence of DP-estimates creates the *D(t)* function of DP-deviations.

The step of the sliding analysis was chosen in the range from 1 to 8 segments (the duration of one segment is about 11 ms). Increasing this parameter coarsens the processing but decreasing leads to the increase of the amount of calculations, so this parameter is compromisely chosen during the optimization.

Let us consider the DP comparison for one step. The comparison process for the word template with *L* length and a certain signal part is presented in Figure 34 by DP-matrix.

Since the words boundaries contained in the input signal are unknown beforehand, so for the testing of the word hypothesis it is necessary to take the part of the signal with some reserve. The length of the part was experimentally chosen *4/3 L*. Thus the size of the DP-matrix equals *L* x *4/3L*. In contrast to common acceptable usage of DP-algorithm the result of

comparison is determinated not as the “right lower” value but as the minimal value in the last column (with the exception of section (*)). The section (*) is the forbidden area. For two-time warping its length equals *1/2 L*, so the interval length, where the minimal distance value is found, equals $\frac{4}{3}L - \frac{1}{2}L = \frac{5}{6}L$. Thus on every step of the sliding analysis the DP-comparison is performed. The length of the signal part is chosen from *1/2 L* to *4/3 L* so that the result of the DP-comparison would be minimal.

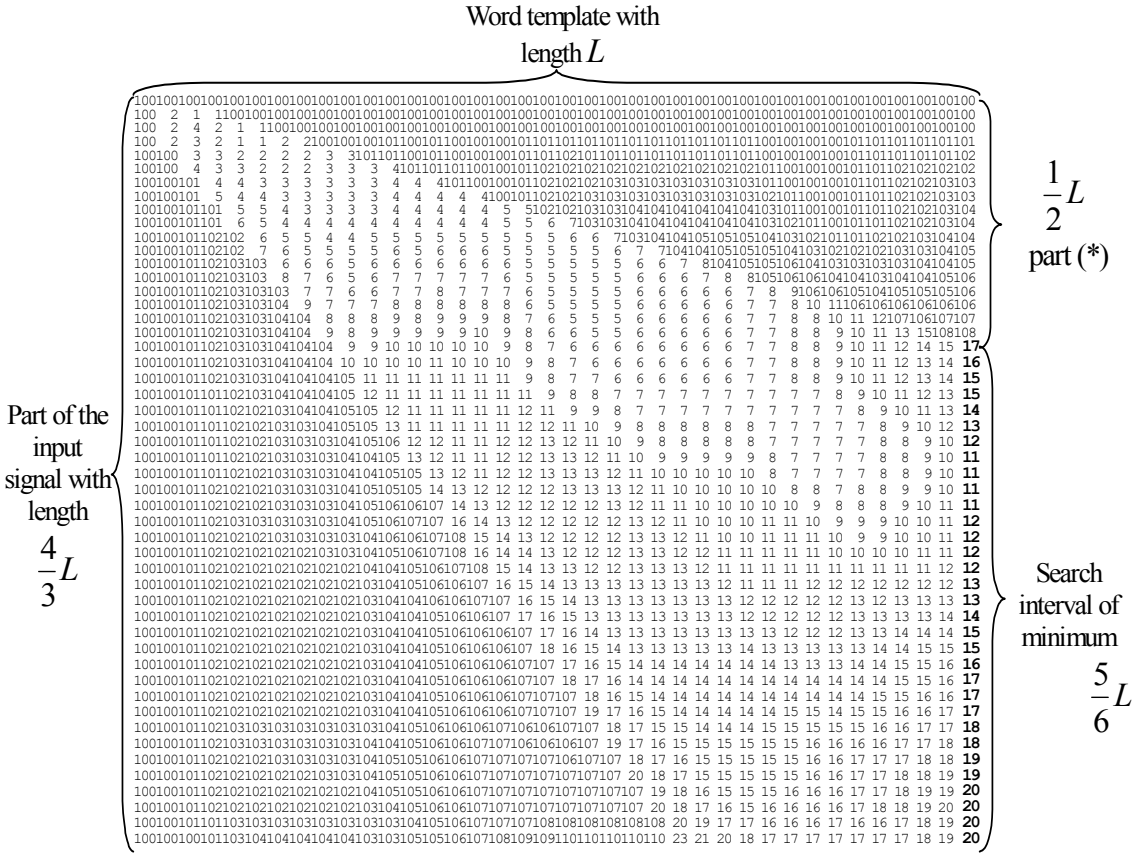


Figure 34. The result minimization of DP-comparison by distance matrix

As the result of the sliding DP-analysis the function of difference between the word template and the signal is obtained. In order to detect the hypotheses of the words appearances we find the local minimum of the *D(t)* difference function. The usual condition is used for the determination of the local minimum. In order to decrease the influence of insignificant changes of *D(t)* function a zone of insensitivity Δ is introduced. The Δ parameter allows to differentiate significant local minimums (supposed word hypotheses) and the insignificant ones. It is clear that it requires the optimization. The search process of local minimums of *D(t)* difference function is shown in Figure 35. Here we can see two local minimums LM_1 , LM_2 and one insignificant minimum.

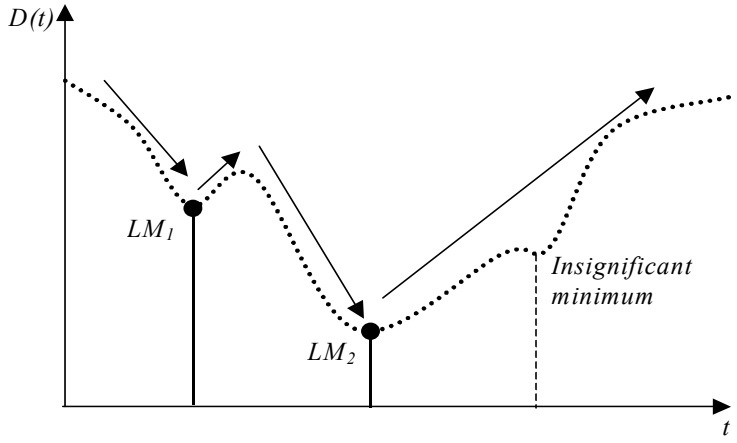


Figure 35. The search of local minimums on *D(t)* difference function

For reducing the number of false local minimums in every analyzer the *P* threshold is used, which characterizes maximally accepted value of DP-difference. If the found local minimum is not higher than *P* parameter then the word hypothesis is set for it and the following parameters are filled: (1) the time moment corresponding to the local minimum; (2) the value of *D(t)* function in this point; (3) the number of the word template, to which the analyzer is adjusted. The increase of the given parameter raises the risk of losing the correct hypothesis but the decrease leads to an unjustified increase of the hypothesis number, so this parameter is determined during the optimization process. Then the data from the word analyzers are joined in the united stream of all word hypotheses.

During the construction of phrases hypotheses all possible combinations of word hypotheses are formed. Their time compatibility is estimated by the fuzzy sets theory. The probability function of the interval duration between word hypotheses has three time sections (Figure 33). The following dependence shows the correlation between these time sections and determines the value of *B* probability function of the time interval between word hypotheses appeared at the time moments t_1 and t_2 accordingly.

$$B(t_1,t_2)=\begin{cases}0,\ t_2-t_1<ZI\cdot L;\\ \dfrac{t_2-(t_1+ZI\cdot L)}{2\cdot ZI\cdot L},\ ZI\cdot L\le t_2-t_1<3\cdot ZI\cdot L;\\ 1,\ otherwise.\end{cases}$$

where *L* is the length of the word template, which appeared at the time moment t_1;

ZI is the parameter, which determines the relative length of the *zero interval*.

In general case the acoustic estimate for the phrase hypothesis equals the multiplication of all the estimates of the acoustic probability of word hypotheses and estimates of the interval probability between them. Such processing allows to suppress false local minimums and avoid false word combinations already at the acoustic processing stage. The number of phrase

hypotheses is chosen by the optimal way so that the correct hypothesis would be in this subset with sufficient likelihood and at the same time the processing complexity would be minimal.

Thus the parameters of the model, which are to be optimized, are the following:

AS – the step of the sliding analysis;

Δ - the zone of insensitivity;

P – the maximally permissible value of DP-difference;

ZI – the relative length of the zero interval.

4.3.3. Optimization and testing of the model. Experimental results

In order to adequately estimate the efficiency of the developed algorithm of continuous speech recognition and optimally adjust the parameters of the model the following criteria were used. *Word recognition accuracy* is used as the main criterion of the optimization. Another criterion is the algorithm complexity, which depends on the number of word hypotheses and the size of the subset of best phrase hypotheses. The optimization task is defined as the search for such values of the model parameters, which minimize the algorithm complexity at the assigned recognition accuracy.

The accuracy of word recognition is determined by comparison of the input word chains with the output ones. It is known that a sequence of continuously pronounced words has been entered into the input. At the output of the algorithm we obtain the best hypothesis of the pronounced phrase in the form of the most probable chain of words from the vocabulary. Then comparing the texts of the pronounced phrase and the recognized word chain we can estimate how many errors of word recognition were made.

The result of the recognition is shown in Table 7 (the length of hypothetical phrases is 3 words). At the first place is the phrase hypothesis, which has 2 errors. The correct hypothesis is placed on the 8th place. The estimate of this hypothesis is not much worse than the estimate of the first hypothesis which allows to choose the correct phrase hypothesis by further high level processing.

Table 7. Recognition example of continuously pronounced words

The best phrase hypotheses with the length of 3 words	Probability
control crew crew	411
control wheel crew	395
control turn_off crew	390
control to_nominal crew	386
control failed crew	381
control crew to_zero	380
control turn_on crew	380
control wheel to_zero	***378***
control crew wheel	378
control turn_off to_zero	375
control wheel wheel	375
control failed to_zero	374
... ...	...

The speed of the recognition algorithm and also the further high-level processing mainly depend on the total number of the word hypotheses, which is obtained at the set output of word analyzers. The detection complexity of the word hypotheses in the continuous speech is proportional to the size of the word vocabulary but the complexity of the construction of phrase hypotheses has exponential dependence on the number of word hypotheses. So it is necessary to control the number of all word hypotheses, which are obtained at the output of the analyzers' set.

During the construction of phrase hypotheses their estimate by time compatibility of input word hypotheses is accomplished. As a result the word chains are obtained both probable and absolutely improbable. In order to accelerate the further high level processing it is necessary to reduce the number of phrase hypotheses obtained at the recognition stage. For this aim it is required to choose the optimal size of the subset of the best phrase hypotheses. It was determined during the series of preliminary experiments so that the correct hypothesis would come into it with the likelihood not less than 90 percents. It must be pointed out that the resulting accuracy of the understanding model is usually higher.

The following databases have been created for testing of the developed algorithm: (1) the vocabulary of word templates; (2) the testing database of continuously pronounced words. Two template vocabularies of 10 digits and 100 words have been created. The test database with continuously pronounced words and template database were recorded by the same speaker. The database, composed of the records of 100 continuously pronounced word sequences made by 3 speakers, was used for estimating the recognition accuracy.

The optimization of the parameters was performed in the following way. At first the initial values of the parameters and limits of their variation were chosen by heuristic method. As the algorithm parameters are principally dependent, the final optimization was performed by the method similar to a complete enumeration of combinations. During the optimization about 1000 parameter combinations were tested, which describe the whole space of their values. As a result of the algorithm optimization the following parameter values were chosen (*AS=6, Δ=60, ZI=4, P=250*).

This optimized model was tested on the following testing material: (1) continuously pronounced digits; (2) the rephrasing database for the demo-version of the model for a voice operated flying object (Table 8). The maximal acceptable length of the hypothetical phrase *L* was *4*.

Table 8. Performance estimation of the continuous speech recognition model

Speaker	Recognition accuracy of continuously pronounced digits, %	Recognition accuracy of words in continuous speech, %	Presence of the right hypothesis in the 10 best ones, %
1	96	74	85
2	93	72	83
3	95	74	86

Moreover the complexity of the model was compared with complete enumeration method and isolated word input (Figure 36). The obtained results allow us to say with certainty that the algorithm complexity is significantly lower than the method of complete enumeration and with the increase of vocabulary size and the length of a hypothetical phrase the unacceptable increase of phrase hypotheses is not observed. On the other hand since the semantic syntactic constraints are not used on the stage of phrase hypotheses construction, the model recognizes both correct phrases and phrases with some inaccuracies that gives the possibility of the further robust speech understanding.

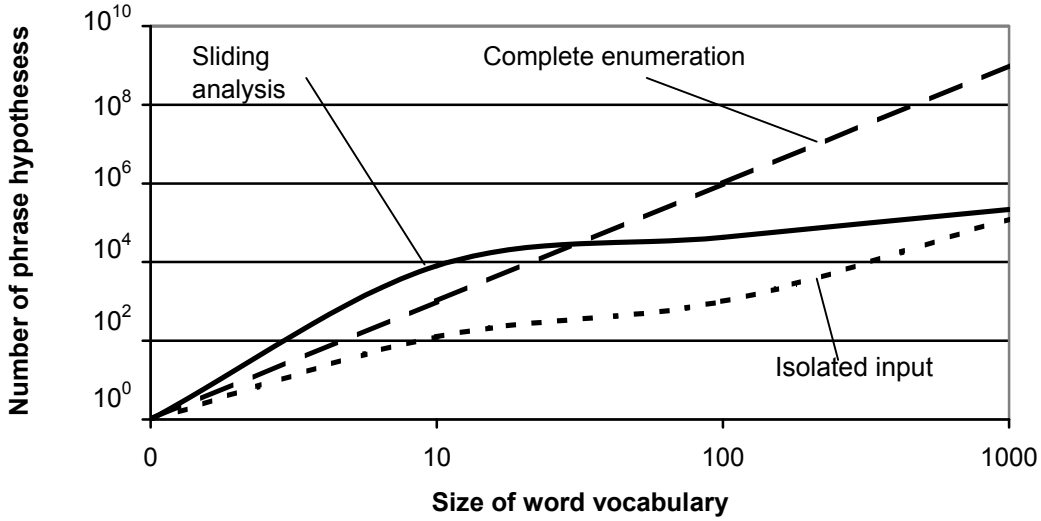


Figure 36. The increase of the amount of phrase hypotheses with the increase of vocabulary size

Introduction of the sliding analysis model into the speech understanding model and their testing is presented below. The speech understanding model was tested by syntactically correct phrases and also the phrases with some deviations. The robustness estimation of the understanding model with respect to errors of the recognizer was obtained.

4.4. Research of the problem of speech recognition/understanding robustness

The robustness problem of speech dialogue systems rose during last years. It interests the researchers and consumers of such systems more and more. It happened because there were many unsuccessful attempts of real applications of speech technologies. It becomes clear, that common parameters such as the vocabulary size and recognition accuracy hardly exhaust the complex notion of the quality of such systems. It is also required to know how a system's performance degrades when it is transferred from an ideal environment of the laboratory to real conditions, when many conditions degrade, unexpected variations of speech signal appear and the speaker changes.

The term “robustness” is now still being used in speech technologies exclusively to show the intention to improve the system’s stability to unfavorable but possible input influences and even to its own errors. At the same time there are not any methods for evaluation of robustness degree. Therefore we cannot compare the robustness of different systems/approaches to diverse impeding factors or to their combination. The creation of objective robustness criteria would allow to exclude subjectivity from the system estimation.

In this part it is shown that the robustness of people speech perception is much better than that of the machine. Some examples of intuitive comparison of several well-known methods for speech understanding are adduced. An attempt is made to offer the definition of robustness degree within the framework of the integral conception.

4.4.1. From robust speech recognition to robust speech understanding

In speech recognition area the sufficient experience has been collected to begin the formalization of the term “robustness”. When we wish to estimate the systems’ ability to resist such factors, as additive noise, variations of the signal level, variation of the microphone position, etc., we can use the experimental way. For this goal we vary these factors and estimate recognition accuracy. So we can detect the size of acceptable impeding factors.

Thus the possibility appears to compare different systems/methods by acoustic robustness. But when we wish to find out the system’s robustness to phonetic incorrectness, the special methods and databases are required.

But things are more difficult with speech understanding. Surely, everything said above for the word recognition robustness is important also for understanding robustness. However, syntax, semantics and pragmatics (especially, situational information) give us great additional resources, which have to be understood and used for the development of new robust methods for speech understanding.

According to definition [99], robustness of language (texts) understanding presupposes the ability to resist diverse grammatical deviations. In automatic speech understanding there may be syntactical errors/variations made by the speaker, or/and word recognition errors made by the system. Thus, to estimate the robustness of understanding, at first we must be able to estimate high-level deviations of input utterances by some quantitative measure. Then we must be able to simulate and collect them in a particular database. At last, it is required to test the understanding accuracy. Of course, the elaboration of some estimation methods and particular test databases will be required.

The speech communication process between people serves as an excellent prototype of future systems for robust speech understanding that is commented below.

4.4.2. Robustness of human’s speech communication

It is useful to remember that speech communication among people possesses striking robustness. This is the opinion of well-known experts in the field of information theory [120]:

"The human language structure is quite complex and allows to conduct a conversation in spite of many noises and distortions. A sensible information transfer can be continued even when a considerable part of syllables is lost. In this sense it is possible to suppose that languages are examples of effective error correcting codes, which proved to be well adapted to people's requirements in the result of evolution."

To compare the human's and machine's speech perception we use speech psychology data [153] and averaged data on machine perception, for example [11,42,117]. Figure 37 shows phrase perception rate obtained by psychoacoustic research (upper curve), and automatic speech understanding accuracy (lower curve).

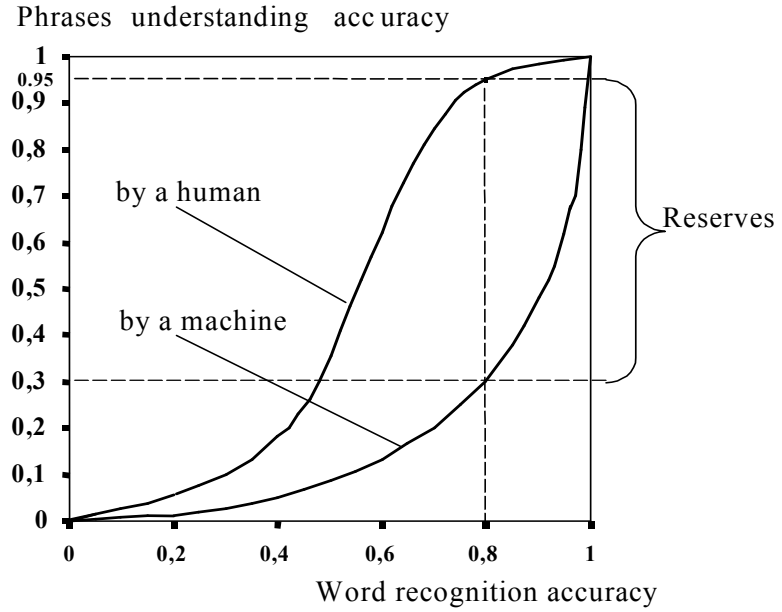


Figure 37. Comparison of robustness for human and machine speech understanding

As the result it can be expressed the important conclusion on robustness of people speech perception: the phrase perception accuracy exceeds 0.95 if word error rate doesn't exceed 20%. In detail these relations are presented in Table 9.

Table 9. Allowed rate of human's word perception for required rate of understanding accuracy

Required rate of understanding accuracy,%	Word perception error, %
95	20
90	28
80	34
70	40

If we accept the usual definition "the system's robustness is the ability to save its performance in unfavorable conditions", so in this case the robustness of the hearing system can

be estimated as top error rating value (20% for instance) at which the system keeps the required level of understanding accuracy (95% for example).

Now let us consider the robustness of machine speech perception [55], which can be estimated by means of Table 10.

Table 10. Allowed rate of machine word recognition for required rate of understanding accuracy

Required rate of understanding accuracy, %	Acceptable error rate of word recognition,%
95	0,5
90	1,0
80	2,0
70	3,0

The comparison of these tables shows that robustness of the automatic speech understanding systems to word recognition errors is very low. For these adduced points the robustness is 40, 28, 17 and 13 times correspondingly lower than that of people.

Of course the criterion offered here is insufficient, because it takes into account only word recognition errors. But there are also wide spectra of diverse kinds of syntactical deviations peculiar to natural speech: word omissions, extra words insertions, word order change, etc. There are not any quantitative estimations of syntactical deviation now, though this phenomenon permanently arises in science and engineering and requires worthy attention of scientists.

Summing up everything said above we should point out that the robustness of human understanding exceeds by far that of the machine in all considered aspects. It should be admitted, that our hearing successfully solves the main problem of artificial intelligence: “how to be exact in respect to inaccuracies”, or in other words “how to get accurate meaning from inaccurate words”. Therefore it is very important to understand the human’s perception nature for the progress in the dialogue with computer.

4.4.3. Recognition of meaning of phrases robust to grammatical deviations of an input message

The developed integral approach to speech understanding task provides required accuracy and robustness to deviations in the pronounced phrase as well as to the word recognizer’s own errors in continuous speech.

Since the main aim of research of integral approaches to speech understanding consists in improvement of robustness (recognition, understanding) it is necessary to try to estimate it. It is clear that robustness must be estimated depending on an impeding factor, which can be estimated objectively. Since not all of them can be quantitatively estimated today it is possible to speak only about some measurable kinds of robustness. Let us consider the main kinds of robustness of the understanding system.

Robustness of the speech understanding system to word level errors. This kind of robustness indicates to the correcting capability of the understanding algorithm to word errors. If during the experiment the noise level is varied and also two parameters are fixed: (1) word recognition errors E_w and (2) phrase understanding errors E_p, then comparing these data it is possible to get the measure of the robustness of understanding: $R = E_w / E_p$

Robustness of the understanding system to grammatical inaccuracies. This is a very important factor with respect to the system's ability to understand the natural language with its inaccuracies. However, it is difficult to get the measure of grammatical inaccuracy. We can speak a lot about different kinds of inaccuracies but it is quite difficult to reduce them to a quantitative value and since there are no methods for estimation of the measure of grammatical inaccuracies of the sentence, therefore it is impossible to estimate this kind of robustness in general. In order to temporarily fill this gap it is possible to use the estimations of understanding accuracy of phrases, which contain diverse types of distortions (extra word insertions, word omissions, interchange of words, etc.), simulated during the experiment. This principle was used in the experiments described below.

The goals of the experiments were: (1) robustness estimation of the understanding model in the speech control task to word recognition errors and (2) understanding accuracy with respect to simulated syntactical deviations. Voices of the three speakers were used during the experiments (Table 11). The size of vocabulary was 100 words; the phrases were taken from the rephrasing database for demo-version of the model for voice operated flying object. The error rate of word recognition of continuous speech was 26,7% on average and error rate of meaning understanding was 1,7%. Thus, the robustness value is 26,7/1,7 = 15,7.

Table 11. Testing the word recognition/understanding of continuous speech

Speaker	Word recognition errors of continuous speech, %	Error rate of meaning understanding, %
1	26	1
2	28	3
3	26	1
Average	26,7	1,7

Thus in the performed experiment the understanding accuracy was not less than 97% even when the level of word recognition errors was high enough. The accuracy of phrase understanding with some types of distortions is presented in Table 12.

Table 12. Understanding of distorted phrases

Type of phrase distortion	Understanding accuracy, %
Omission of a separate word/words	97
Insertion of an extra word	91
Interchange of words	98

This result exceeds all known data about voice control models. It can be explained by the advantages of the totally integral approach and especially by the influence of the situational information.

4.5. The analysis of the approaches for the adaptation of the system to a speaker and acoustic environment

The great difficulty of speech recognition is a large variance of voices of users. These differences depend on the gender, the individual form of the vocal tract (which affects the pitch and the timbre of the voice), dialects, individual manner of sound and intonation formation, emotional state, etc.

At present great efforts are being spent to overcome these difficulties, however the achieved results are insufficient. The investigations are going on in two directions: (1) the search for invariant features and models, which allow to get the similar feature description of words, pronounced by different speakers and (2) the development of methods for the system adaptation to a new voice/environment.

The first direction (invariance to the speaker) has limited abilities and is used in the applications, where it is impossible to employ any kind of adaptation because of exploitation conditions, for instance, in response telephone systems. The limitation of this direction is confirmed by the fact that humans contacting with a lot of people (cashiers, teachers and others) very often have to re-ask interlocutors. So even a human's hearing system is not able to solve this problem absolutely exactly.

On the contrary, the second direction has great reserves; it is confirmed by a wonderfully fast human adaptation to new voices, speech defects, noise, etc. It is a very important task to elaborate these natural mechanisms and apply them in automatic speech recognition devices.

Depending on application conditions, three types of adaptation techniques can be used: a batch type (based on off-line processing of great amounts of training data), incremental type (with gradual self improvement) and an instantaneous one (performed during a short dialogue). The first approach is used in systems, intended for long work with the concrete user [36]. Such systems have the largest increase of the recognition quality, but their negative side is the great amount of preliminary data, which are required for the adaptation process, and therefore the off-line mode is typical here. In case of incremental adaptation recognition quality increases continuously and the adaptation process is accomplished simultaneously with the direct use of a system. Since the preliminary procedure of the adaptation material accumulation is excluded here [109,118]. The instantaneous adaptation is used for such kind of applications, where the term of using a system may be very short, for instance, question-answering systems. The speech to be recognized may consist of only a few words and therefore the adaptation algorithm must work as rapidly as possible. Such models can be based on the modification of the space of speech feature clusters during the recognition of a few words [28,76]. The instantaneous

adaptation is the most interesting category because it tries to simulate the human's perception peculiarity.

Depending on the degree of a human's control of the adaptation process the approaches are divided into supervised and unsupervised [37,78,80]. During the supervised adaptation the speaker reads the prompted text and also corrects the recognizer's mistakes. In the unsupervised mode the system does not require such operations. But the increasing degree of the recognition quality is potentially less.

There are several approaches to the speaker adaptation. At the beginning the methods of spectra signal transformation appeared [13]. Here the normalization of the speech spectrum is reached by linear displacement on the Bark scale. The experiment with speech synthesis has shown that the most important variable in perception is the location of formant frequencies [51]. So methods, which reduce the variance in their location, reduce differences between speaker voices. Using such model allows to reduce the difference between speakers to 15.5-17.0 % [111]. Another approach to reduce inter-speaker variance is the normalization of the vocal tract configuration whose main difference is only in length. It allows to reduce the error rate to 51.8 % [82].

Another popular approach to spectral transformation is based on vector quantization (VQ) technique. The key idea of this method is to establish a correspondence between pairs of typical spectra from two speakers based on their occurrence in the same context speech. As a result the VQ-based spectral transformation maps an arbitrary spectrum of one speaker to another spectrum that is a member of a finite collection of typical spectra of another talker [91].

One more approach to the speaker adaptation is a speaker clustering [49,107]. The idea is to cluster the training speaker into a number subgroups and preliminary train a complete speech recognizer for each group. When the system is used test speakers are classified to the nearest cluster and the corresponding recognizer is used for decoding their speech. The effect of this approach is limited. At the top level it is useful to distinguish between male and female speakers but further divisions are difficult because individual speakers are too different to be effectively clustered. Furthermore, collecting enough training data quickly becomes a problem when many speaker clusters are used.

The wide class of methods using correlation of speech acoustic states based on hidden Markov models (HMM) must be mentioned [16,43,125]. The principles of maximum likelihood linear regression (MLLR) and maximum a posterior (MAP) allow to compute a set of transformations that will reduce the mismatch between an initial model set and the adaptation data. More specifically such model of the adaptation technique estimates a set of linear transformations for the mean and variance parameters of a Gaussian mixture HMM system. The effect of these transformations is to shift the component means and alter the variances in the initial system so that each state in the HMM system is more likely to generate the adaptation data.

During the project the original approach to the speaker's adaptation has been developed. The adaptation can be performed in two modes: (1) the substitution of acoustical states by using usual VQ and (2) the substitution of the segments of signals, corresponding to quasi-allophones. The second mode is more effective. It allows to compensate voice differences, as well as acoustic channel differences/changes [93]. In addition to well-known methods of spectrum normalisation the method taking into consideration the overtone attenuation speed in the vocal tract has been proposed. This parameter changes greatly with different speakers but it is not directly observed in speech spectrum.

A problem of the acoustic environment arises when speech technologies for the control of moving objects are used. Noises usually lead to the degradation of speech recognition quality. In order to reduce the influence of external noise on recognition quality differential microphones, a special communication helmet with noise insulation and other approaches are used. Also robust speech understanding methods are radical issues for compensation of this factor.

The speaker independence problem has two sides. On the one hand, it would seem very good if the system could understand any voice. But, on the other hand, there are a lot of applications, which need a confidential access to a system, for example, in the control of an aircraft, a ship, a banking terminal, etc. Moreover, the performance of the systems adapted to a concrete voice is higher than by using the speaker independent approach. Also it must be added, that the adaptation time for systems with average size vocabulary (several hundreds of words) is not very long. So taking into account everything said above, the individual adaptation fits well for our task.

4.6. Development of flexible structure of the speech databases for simple adaptation to modifications of the object work logic

Any experimental system like this demands numerous adjustments of initial data during development and debugging of a system. At that a user of the system must adjust all the databases and make it in the certain order. To make this work more operative, the especial mechanism, which allows to work with the speech databases in interactive mode (i.e. to automate laborious process of the initial data modification), has been developed.

4.6.1. Organization of the databases of the speech understanding model

The speech understanding model is based on four databases: situational, lexical, associative, and acoustic. Below the structures of these databases are described.

The situational database. This database serves for the description of the complete state diagram of the control object. It is used for estimation of the pragmatic correspondence degree of an input phrase to a current situation. The database contains the list of all possible situations and descriptions of the fragments, which determine possible transitions from the current situation to the following ones. The description of each fragment contains:

- A subset of possible transitions from the current situation to the following ones;
- Commands for the transitions to the following situations. We use the subset of the alternative phrases (paraphrases), since uttered phrase may be involuntarily modified by a speaker;
- Semantic weights of words/syntagmas for each alternative phrase.

General structure of this database is presented in Table 13. The fields with numbers 2, 4, 6 and 9 must be filled in by an expert. All the other fields are automatically filled in on basis of based entered information.

Table 13. The structure of the situational database

Situation №	Description of situation							
	Situation name	Number of transitions	Description of transition to the following situation					
			Following situation №	Number of para-phrases	Description of equivalent phrases			
					Text of para-phrase	Array of syntagma indexes	Number of syntagmas	Weights of syntagmas
1	2	3	4	5	6	7	8	9

A more detailed specification of the database structure is presented in Tables 14-16.

Table 14. Description of situations

Field name	Type	Field description
Title	Textual	Situation name
MaxTransitions	Numerical	Number of transitions to other situations
SituationTransitionDescription	Array	Array of the transition descriptions

Table 15. Description of transitions

Field name	Type	Field description
TransitionNumber	Numerical	Number of the following situation
MaxPhrase	Numerical	Number of paraphrases
PhraseDescription	Array	Array of descriptions of alternative phrases

Table 16. Description of alternative phrases

Field name	Type	Field description
SyntagmNum	Numerical	Number of syntagmas in a phrase
SyntagmList	Array of numbers	Array of syntagma indexes, composing a phrase
WeightList	Array of numbers	Array of syntagma weights
SyntagmTitleList	Textual	Text of paraphrases

Syntagma vocabulary. The vocabulary is the list of all existing syntagmas of the speech understanding model, where each syntagma is associated with the numerical index.

Associative database. This database contains compatibility estimations of all ordered pairs of syntagmas of the vocabulary. To each pair of syntagmas the associative estimation by 4-score scale is given:

4 – excellent compatibility;

3 – good compatibility;

2 – satisfactory compatibility;

1 – bad compatibility;

The structure of the associative database is presented in Tables 17-18.

Table 17. Description of syntagma associations

Field name	Type	Field description
AssociationsCount	numerical	Number of syntagma associations
EstimationsDescription	array	Array of descriptions of the associative estimations

Table 18. Description of associative estimations

Field name	Type	Field description
SyntagmIndex	numerical	Index of syntagma with which the association is set
Estimation	numerical	Associative estimation

Acoustic database. This database contains the parametric representation of all words/syntagmas by chain of cluster numbers. This database consists of the fragments represented in Table 19.

Table 19. Description of acoustic database

Syntagma number	Unquantized feature vectors	Chain of cluster numbers
1	2	3

4.6.2. Adaptation of the databases to modifications of work logic of the control object

For automation of joint database adjustment the module of integral adaptation of databases to modifications of object work logic has been developed.

Using this mechanism the following operations with speech databases can be made:

1) Situational database adjustment:
   a) Addition of a new situation name;
   b) Removal of the existing situation (with all attributes);
   c) Alteration of the situation description;

d) Addition of a new transition from situation to situation;

e) Transition removal (with all transition commands);

f) Addition of a new transition command from situation to situation;

g) Removal of transition command;

h) Transition command change (including weights of command syntagmas);

2) Adjustment of associative estimations for any pairs of syntagmas;

3) Acoustic database adjustment:

a) The creation of the acoustic database for a new user;

b) The alteration of the acoustic templates for a certain user.

In Figure 38 the interaction between the modules of database adaptation and the speech understanding at the stage of debugging and testing is presented.

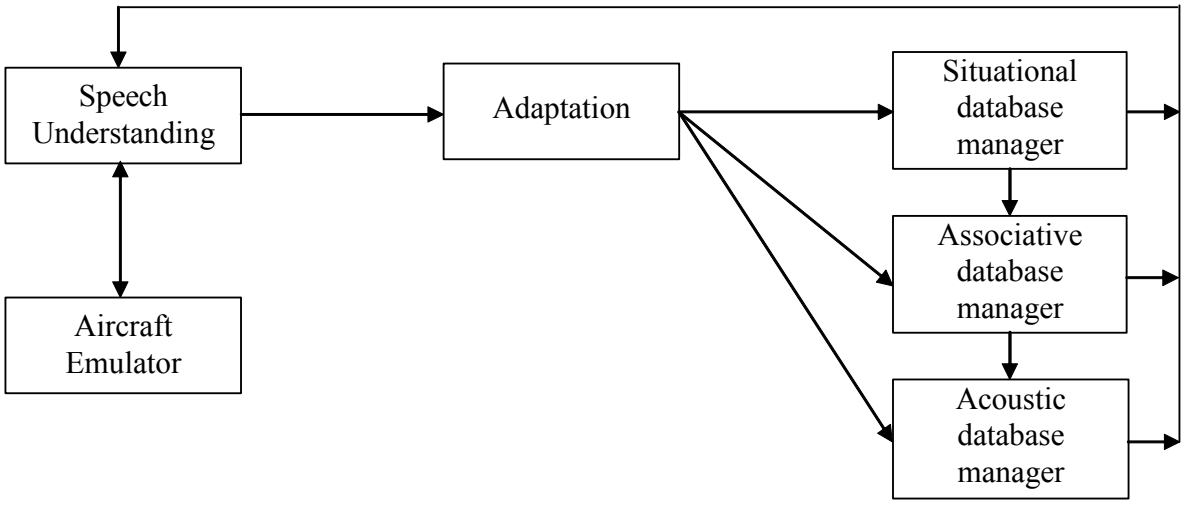


Figure 38. Interaction between the modules of database adaptation and the speech understanding

The order of adaptation of the speech databases is shown in Figure 39. This process is divided into two phases taking into account the collaboration of specialists from different areas. Accordingly the databases are divided into two categories: (1) situational, associative databases and syntagma vocabulary; (2) acoustic information (acoustic templates of the syntagmas). The creation and modification of the databases is conducted sequentially in two phases.

At the first phase the expert enters the initial data about the control object (situational model of the control object). After that the module creates the syntagma vocabulary, parsing the texts of input phrases. The creation process can be performed in three modes: manual, automatic and mixed. In case of using the manual mode the expert fills in all cells of the matrix, using a 4-score scale of estimates. In automatic mode the adaptation module sequentially parses all phrases, which contain in situational database and recognizes the presence of the ordered

pairs of syntagmas. For the detected pairs the estimate “4” is entered in the matrix; for the inverted pairs is “3”. For all other pairs of syntagmas the estimate “1” is given.

At the second phase a user enters all syntagmas/words of the syntagma vocabulary using a microphone. Thus the module creates the acoustic templates of these syntagmas.

The information entered at the second phase of database adaptation, depends on the user voice only and here the participation of the experts and the developers is not required. On the contrary, the first phase is completely dedicated to accumulation of high-level information about the object. Here the collaboration of developers and experts of applied area is necessary. The software, which realizes the integral adaptation of databases, is described in the following sections.

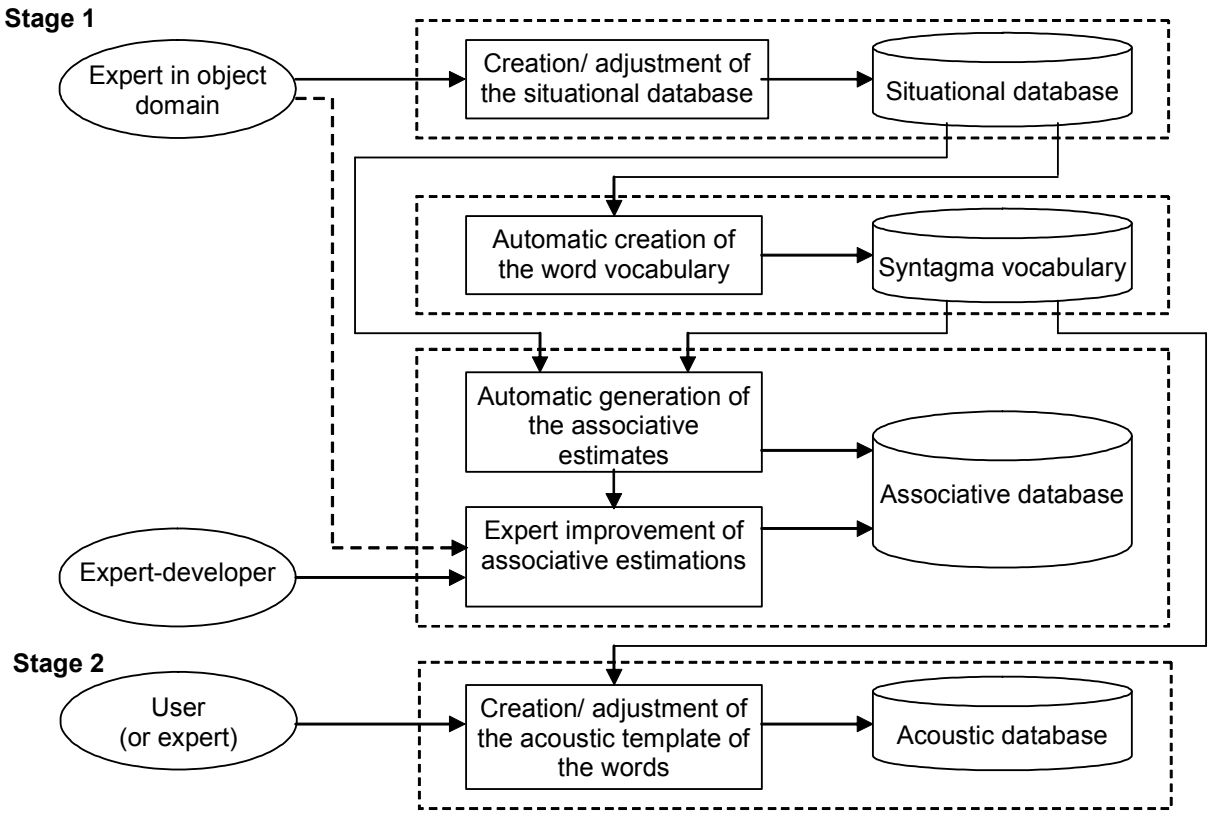


Figure 39. Integral adaptation of the speech databases to modifications of object work logic

Usage of the developed mechanism for adaptation of databases at the stage of debugging and testing of the understanding model gives some advantages:

- The required flexibility of the databases structure.
- Friendly interface by means of prompts and recommendations to a user.
- The required efficiency owing to the unification of the understanding model with the module of integral adaptation of databases in one complex.
- The significant reduction time for databases adjustment.

4.7. Modification of the base model of the integral understanding in order to provide the continuous speech input and adaptation to applied task

During the research the following modules were created (1) the integral adaptation module, which provides the inter-coordinated adjustment of all databases of the understanding model as well as the portability of the understanding model to new applied tasks; (2) the module of continuous speech recognition robust to grammatical deviations based on the sliding analysis of a speech signal and a posteriori phrase hypothezation. The developed modules have been included into the base version of integral understanding model. As a result the united software complex has been created. It provides robust understanding of continuous speech as well as the portability to new applied tasks due to integral adaptation.

The structure of the adaptive model of robust understanding of continuous speech is shown in Figure 40. The model contains three levels of information processing: (1) acoustical-lexical; (2) semantic-syntactic; (3) situational-pragmatic.

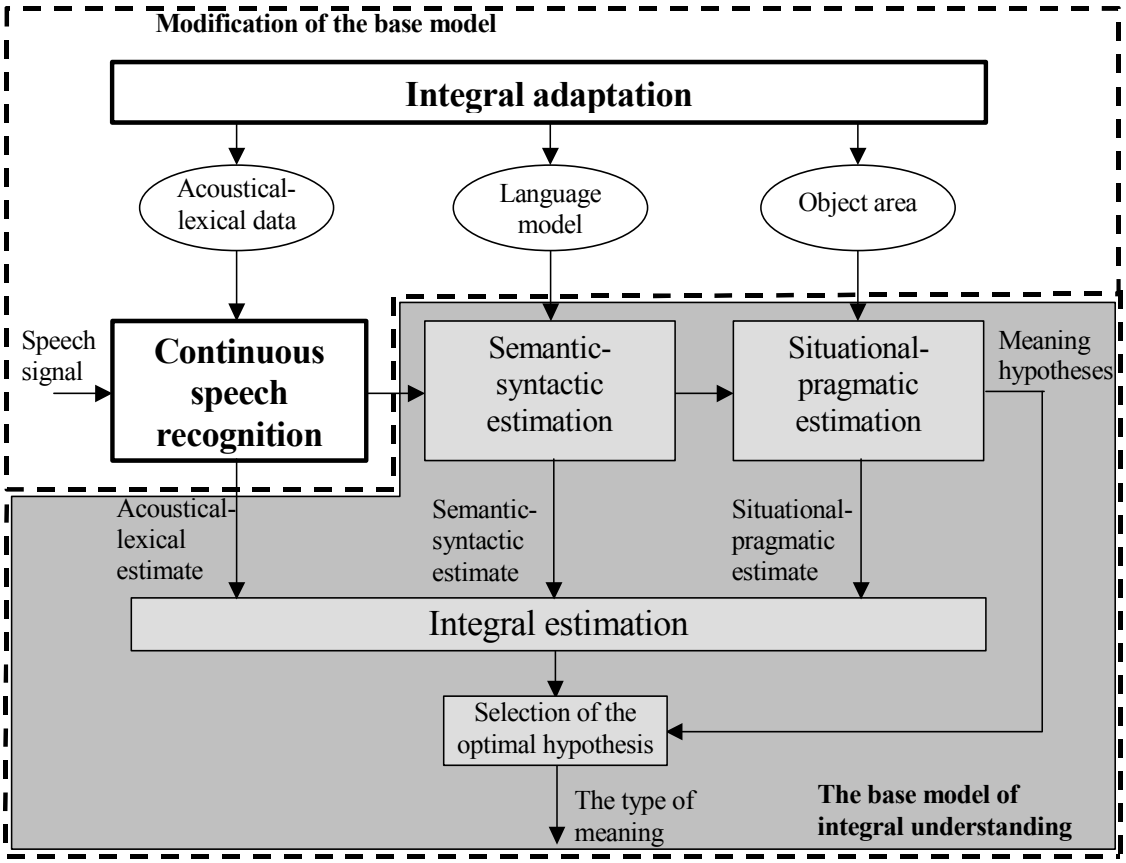


Figure 40. The structure of the adaptive model of robust understanding of continuous speech

Understanding of a spoken message is carried out due to integral interaction of all partial processing levels. Every recognition level gives a set of alternative decisions with corresponding estimations. The number of possible indeterminacies and errors, which were

entered in the current level (or appeared on it), decreases during integral processing. The essence of integral processing is based on the estimation of the input signal by criteria of corresponding knowledge. The integral estimation is calculated on the basis of separate estimates and the final decision is made by the minimum of integral deviation. It makes the model robust to probable inaccuracies in the pronounced phrase and distinguishes it advantageously from the commonly accepted conception of sequential parsing.

The developed methods of the robust speech endpoint detection, feature extraction, continuous speech recognition and the integral structure of processing provide the robustness to various distorting factors (acoustic-phonetic and grammatical deviations in the pronounced phrases, etc.), that makes the character of interaction between a user and the system more natural. The adaptability of the understanding model is provided by the functions of the integral adaptation module. This module realizes the inter-coordinated adjustment of all databases of the understanding model.

In contrast to the typical understanding model (Figure 2 left) the developed model has the following advantages (Table 20):

Table 20. Comparative analysis of understanding models

Criterion	Typical model	Developed model
Accuracy	The consequent processing does not provide efficient accuracy	Higher accuracy is achieved by integral approach
Robustness	The model is not capable of processing the input phrase with grammatical deviations	Robustness is provided by the robust speech encoding method, new approach to continuous speech recognition and the integral structure of processing
Adaptability	Adaptability is provided on the acoustic level only	Adaptability on all the levels is provided by integral adaptation

Thus the model of continuous speech understanding is obtained, which can be applied in diverse human-machine systems with the speech interface [95].

4.8. Situational aspect in the tasks of speech understanding

In this section we would like to consider the problem of using the pragmatics (in the first place, situational information) more widely. Up to now the specialists in the area of the development of speech understanding systems mainly used the linguistic knowledge of speech as an abstract system beyond its communicative function. For the further improvement of the quality of understanding systems it is not enough to improve the speech recognition module. Huge resources are contained in situational information and their effective use combined with lingual and acoustic knowledge is the basic subject of investigations of perspective speech

systems. Therefore in this section we make an attempt for a wider analysis of the pragmatics and first of all, the situational aspect.

4.8.1. Situational analysis: philosophical and psychological aspects

Up to the middle of the XXth century in the majority of language investigations the researches did not pay due attention to the communicative function of language and so the results of these investigations cannot be effectively included into the model of *speech activity*.

Many works on automatic understanding of language and speech are mainly based on fundamental linguistic theories with obvious grammatical accent. Speech is considered as an abstract system regardless of its communicative function. The natural language understanding models, which were obtained during these investigations, are deprived of a very important component such as the knowledge about situational information. This information as it is well known to the philosophers [135] and the psychologists [142] sufficiently reduces ambiguity and helps to understand language and speech quite completely.

In order to get over the gap between the purely linguistic and the communicative side of the speech process it is necessary to learn against the investigations in psychology of speech activity. This scientific direction deals with studying human speech activity, human's thinking and interaction with the real world. Speech is considered within the framework of the communicative activity of the human, taking into account causal-even and spatial-temporary connections. It is important for us to extract basic regularities and peculiarities of speech activity to apply them for creation of human-computer systems.

In the course of the development of psychology of activity it became clear that each human action, including speech is connected with striving to change the surrounding world [5]. Different theories of activity were developed, including the theory of speech activity. Therefore the reconsideration of the scientific speech paradigm and the inclusion of a human as an active user of speech took place. Consequently new linguistic theories appeared. One of them is the Speech Acts Theory [5]. Here the basic unit of speech is "the speech act" (not an isolated utterance). Speech act unites a single intention, a completed minimum segment of speech and an assumed result. The definition of speech as the means for the aim achievement is the most important for speech understanding researchers. In this case the meaning of an utterance is determined by the speech intention of the speaker.

In order to recognize the speech intention it is necessary to have the information about the conditions in which this intention has appeared. The Speech Communication Theory says about it: "the initial moment of any speech action is the speech situation, i.e. such circumstances which stimulate a human to a speech act"[143]. Therefore we can conclude that the situation plays the primary role in the formation of the speech intention and it is necessary to consider it in more detail.

Considering a situation we can apply definitions from different areas of knowledge: psychology, philosophy, control theory, etc. These differences are conditioned by the specificity

of each branch of science. We are interested in those definitions, which point to the role of a situation in human activity and determine the dependences between the situation and the meaning of a speech action.

In situational psychology the situation is considered “as that general thing, which appears between psycho-physical world and a subject which is in an informational connection as spatial-temporary and psycho-physical now, belonging to subjective and objective worlds in equal degree.”[142].

This definition agrees with the definition of the situation in the control theory, as it is based on the knowledge of psychology of human behavior while deciding of a control tasks [154]: “Let us call the current situation the total of information about the structure of the control object and its functioning at a given moment.”

The definitions mentioned above do not completely point out how the situational information is presented in human consciousness. Moreover, the situational psychology proposes its own conception of situational information. In accordance to this conception, the information about the surrounding word is stored as situational phenomena in human consciousness. In general, the situational phenomenon is a group of acquired skills. These phenomena are in latent state and they are actualized at the relevant moment, when a human finds himself in relevant to the phenomenon conditions. Thus, human consciousness leans against the situational world image, which contains the situational phenomena acquired in the process of vital activity.

The problem of reflection of reality in the human consciousness is also studied in psycholinguistics. The attempts of “modeling situational interaction between the human and the world in the direction of creation of “psycholinguistics of activity” or “psycholinguistics of active interaction” have been already made”[146]. For instance, within the framework of cognitive psycholinguistics the theory of scripts and scenes is developing [103,163].

From what was said above it is possible to conclude that situational analysis represents a situational world image as a set of situations, which is realized at certain relevant moments. However, it is far from being everything, which must be said about the situation and its role in meaningful perception of an utterance. However here it is important to take into account logical-temporary connections between situations and the connection between situational and lingual aspects.

The complexity of meaningful perception is conditioned by polysemy and homonymy of words, which compose an utterance. This ambiguity is significantly reduced by the concrete communicative situation and speech context.

The deepest analysis of human perception in the framework of artificial intelligence problems was made by American philosopher H. Dreyfus [135]. In his opinion a human perceive the world as a context of his pragmatic activity.

It is necessary to understand that the human speech perception is based on situation frameworks, formed by the context. The importance of creation of the situational framework is

emphasized by the speech researches [128]. They say that the creation of the contextual framework, which is wide enough, allows to develop programs, which are able to recognize language images.

Because of the evident lack of scientific knowledge about pragmatic information, approaches to formalization and use of situational information are not developed enough. In the obvious form the situational information cannot be found in books, handbooks, dictionaries, etc. Nevertheless we can try to outline the structure of the situational framework of a certain applied area or a theme by analogy with the situational world image reflected in the human consciousness.

4.8.2. Generalized form of presentation of situational information

Our hypothesis consists in the following: in any kind of speech activity the same mechanism of situational processing is used. In accordance with the conception of situational phenomena and taking into account contextual relations between them, we can propose the following:

- Situational world image is presented as a constantly supplemented set of situations. A theme or applied area can be reflected by a set of corresponding situations.
- The situational framework can be obtained by superposition of contextual relations between corresponding situations.
- Each transition from situation to situation reflects the essence of a speech intention, namely the meaning of an utterance. The meaning of the sentence is defined according to Melchuk [148] as “the invariant of all synonymous transformation”. This definition is based on the assertion that "the possession of meaning" is realized by the speaker as the ability to express the same idea in various forms, and it is realized by the listener in the comprehension of semantic identity or similarity of formally different sentences. Therefore, the transition from situation to situation is reflected by the set of possible and required rephrasings of a speech utterance.

We propose the generalized presentation of situational information in the form of the situational database as the oriented graph. Its arches are transitions from situation to situation; each arch is connected with the subset of equivalent sentences, which define a certain speech intention or a concrete command. A fragment of the situational diagram is shown in Figure 41. Any situational diagram can be broken into such fragments.

In our works [56] the model of the applied area for restricted class of tasks, connected with the control of technical objects (a car, a plane, a robot, etc.) is proposed. But there are no techniques of creation of situational databases for other intellectual applications of speech technologies. The method of the creation of the situational database based on text analysis is presented below.

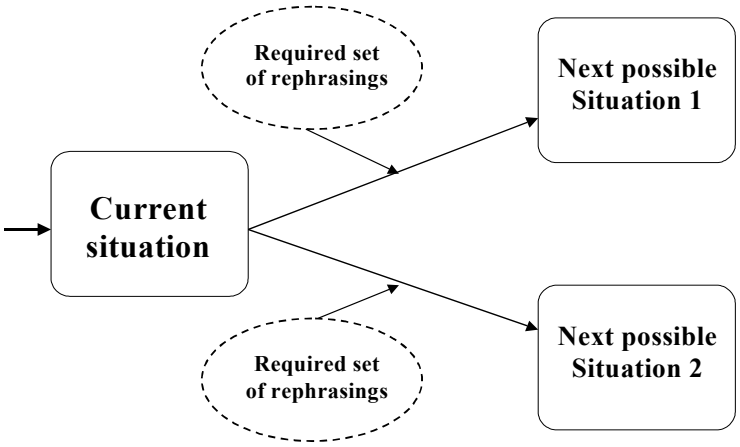


Figure 41. Fragment of the situational diagram

4.8.3. Expert approach to situational analysis

The basic problem of situational analysis is the creation of situational databases. These databases are not formally fixed in any bulletin or handbook, even in such source of speech databases as ELRA (European Language Resources Association). The existent idea of situational world image is not complete. On the one hand, it does not reflect interconnections between situations and, on the other hand, it does not reflect the connection of the situational and the speech aspects.

Speaking about the creation of situational databases it is possible to divide them into two significantly different classes: (1) numerous intellectual applications such as spoken language translation, voice question-answering systems, systems for knowledge and skills teaching, etc.; (2) the method for creation of the situational databases for technical objects control systems.

Let us consider the model of the extraction of situational information in the process of text analysis. Let the analyzed sentence consists of 3 words as shown in Figure 42. Each word can separately stimulate in the reader's consciousness (subconsciousness) a certain subset of life situations, (for example, SS1, SS2, SS3). The word sequence of the phrase can initialize a significantly more reduced subset of situations PSS in the reader's mind (hatched segment). Using the mechanism of the set theory we can present this subset as a result of the intersection of the subsets: $PSS = SS1 \bigcap SS2 \bigcap SS3$.

The result of such processing cannot be an empty set, since the sensible sentence must be connected with a certain life situation. However an ambiguity is possible when PSS contains 2, 3 and more hypothetical situations. Apparently, in the understanding process, a further use of this information can be realized owing to a wider context or due to the use of quantitative estimations and the optimal hypothesis selection.

It is clear that at the first stage such models can be obtained by means of a manual text processing by experts. In a certain restricted text the expert defines the subset of situations, which represents the meaning of the text adequately. Then he segments text into fragments

corresponding to these situations. Then the construction of subsets SS1, SS2, etc., as well as PSS can be entrusted to a program. Such way may seem very consuming, however, if there are no other ways then sooner or later the methods for automatic or semiautomatic solution of such problems will be found. The large world experience in creation of complicated speech systems proves it.

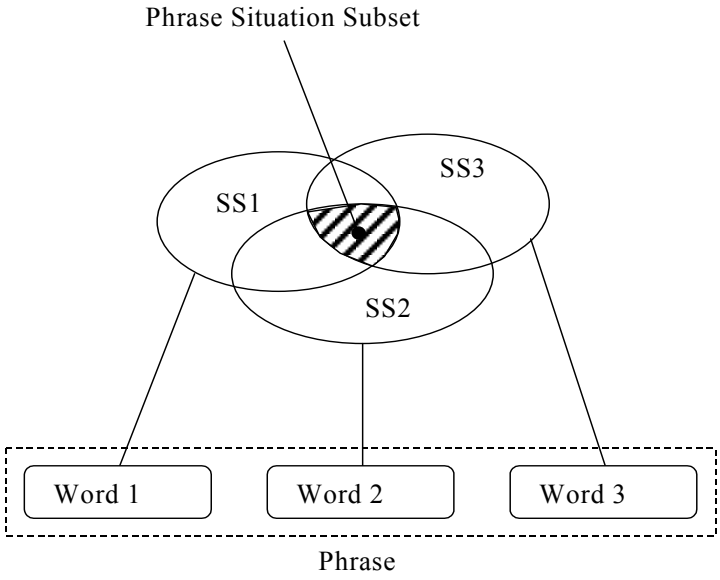


Figure 42. Situation analysis of the pronounced phrase

5. The developed model of the voice operated flying object

In this section the developed research prototype of speech understanding module and the demonstration model voice operated flying object, which includes the aircraft emulator and the speech understanding module are described. The software classes for the speech recording, recognition and understanding are considered more detail. Then the manuals for the research prototype, the demonstration model and additional programs for databases adjustment are adduced.

5.1. Description of the developed software for speech understanding

During the project the software complex has been developed, which includes: (1) the research prototype of the model for voice control; (2) the demonstration version of voice operated flying object, which includes the speech understanding module and the aircraft emulator; (3) auxiliary programs for creation and adjustment of the speech databases. The software was realized on the C++ language using the principles of object-oriented programming (classes, inheritance, encapsulation, etc.). The developed classes, their methods, input/output data, inheritance ties and also the complex of methods and means used for debugging and testing the software are described below.

5.1.1. Main classes of the developed software

The developed software allows to accomplish the speech signal processing control of the emulator of flying object. The structure of base classes and inheritance ties between them are presented in Figure 43. The inheritance relations are shown by solid line. The dotted lines is used for instances of classes.

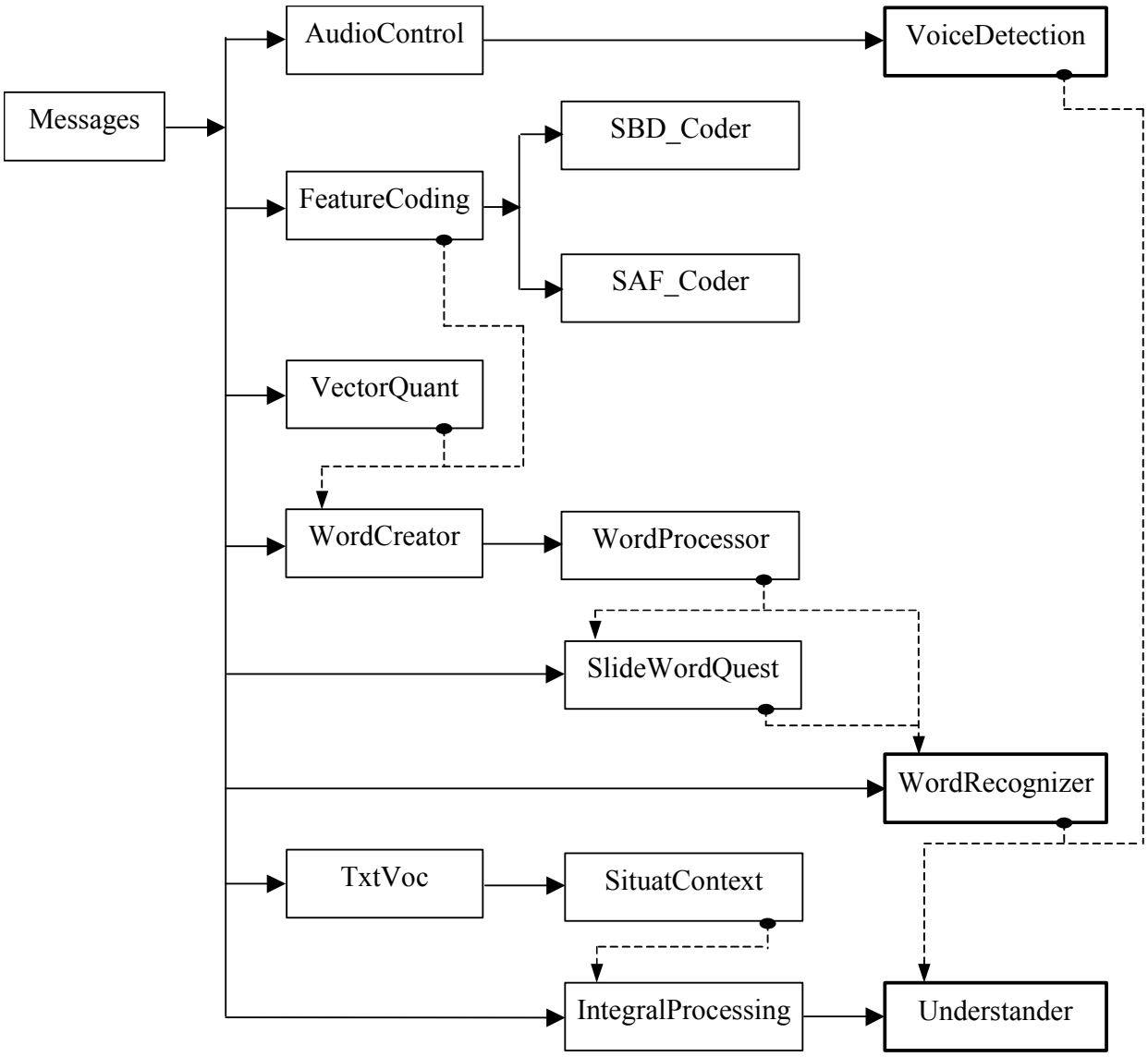


Figure 43. The structure of classes of the integral understanding model

"**Messages**" – the class displays the messages about errors on screen (for instance, "Memory is not enough", "Wav file of response phrase is not found", etc) and it is inherited by all the other classes. The method vDisplayMessage displays the window with the corresponding message.

"**AudioControl**" – the class provides low-level work with audio channel. Methods of this class realize the initialization of audio channel, playing signal, recording signal and segmental recording. The most important methods of the class are:

bEnableRecordingAudioChannel – realizes the initialization of the audio channel. Its parameters are:

ulMaxRecordTime – maximum recording time;

ulFrequency – frequency of the signal discretization during recording;

ulSegmentSize – length of the speech segment;

ulPlayback_mode – frequency of the signal discretization for playing back the recorded signal (if it is zero, then the playing is not accomplished);

The method returns true/false depending on result of the performance.

ResetRecordingAudioChannel – freeing audio channel and returning to the initial state. The method returns value true/false depending on result of execution.

bOneSegmentSignalRecording – realizes recording one speech segment from the microphone (the length of speech segment is set at the initialization of audio channel). The method returns value true/false depending on result of the execution;

vDisableRecordingAudioChannel – realizes the turning off the recording mode. The method does not have input and output data.

vPlayVoice – realizes playing the audio data located in the memory at the pointed address. Parameters are:

pSndPtr – the address of the buffer with audio data;

ulSndLen – the size of the audio buffer,

ulSpeed – the frequency of the signal discretization during playing.

"**VoiceDetection**" – is a child of the "AudioControl" class and provides detection of word/phrase boundaries in the input signal. It accomplishes the recording of the signal and selection of the useful speech signal. The class has some constants describing stochastic characteristics of speech. The main methods of the class are:

shVADRecording – accomplishes the recording and selection of speech signal. Parameters are:

ulMaxRecordTime – maximum recording time;

ulFrequency – frequency of the signal discretization during recording;

ulSegmentSize – length of the speech segment;

ulPlayback_mode – frequency of the signal discretization during playing back the recorded signal (if it is zero, then the playing is not accomplished);

bOneWordRecord – flag allowing to stop the recording after detection of first word.

The method returns the number of the recorded speech segments.

CodingRecordBound – realizes the recording one segment of signal and categorizes speech/silence. Parameters are:

bAcoustWork – flag allowing to stop the recording after detection of the first word in signal.

The method returns value true/false depending on result of the execution.

vCalcNoiseLevel – accomplishes the adjustment of parameters of the speech endpoint detection algorithm to the acoustic environment. The method does not have input and output data.

vSignalAnalysis – determines primary parameters of the recorded signal (duration, amplitude, etc). The parameter is:

bVolLimit – flag allowing limitation of the signal amplitude.

"**FeatureCoding**" is virtual class. The main methods for the parametric representation of speech signal are described in this class. It contains two virtual methods:

shComputeFeatureVector – method for the parametric representation of speech signal. The sequence of the digital samples is transformed into the sequence of feature vectors. Parameters are:

pBuf – pointer to the buffer, where data of the recorded signal are located;

lBufLen – size of the buffer;

pVectorBuffer – memory address, where the resulting feature vector is stored.

The method returns the number of the processed speech segments, i.e. the number of the feature vectors, which was obtained during processing of the speech signal.

lGetSegmentNumber – method calculates the number of the feature vectors using length of the speech signal. In the input: the number of the ADC samples, in the output: the number of the feature vectors.

"**SBD_Coder**" – is child of the "FeatureCoding" class and provides the parametric signal representation by spectral-difference features method. The class realizes the following algorithms: Fast Fourier Transformation, calculation of the bandpass spectrum of speech segment, calculation of the spectral-difference features. The main methods of the class (shComputeFeatureVector, lGetSegmentNumber) are described in the parent class "FeatureCoding".

"**SAF_Coder**" – is child of the "FeatureCoding" and provides parametric signal representation by the method based on sign autocorrelation function. The class realizes the following algorithms: calculation of the coefficients of zero crossing of signal, calculation of the derivative of signal, calculation of the sign autocorrelation coefficients and calculation of periods duration distribution of constant signs signal. The main methods of the class (shComputeFeatureVector, lGetSegmentNumber) are described in the parent class "FeatureCoding".

"**VectQuant**" – realizes the technique of the vector quantization and dynamic programming. The base methods connected with the clusterization are:

bInitClustersMatrix – initialization of the databases required for realization of the clusterization procedures;

chVectorQuantization – realizes the clusterization procedure (finds the nearest etalon vector to the input vector). It returns the number of the nearest etalon vector.

vVectorChainQuantization – accomplishes transformation of the sequence of vectors into the sequence of the corresponding numbers of clusters.

The methods for the dynamic programming (DP) are:

shDP_optimal –is DP method, which allows two-time warping. It works with quantized signals. It returns DP-deviation, which is characterized by the degree of difference between the compared signals.

shDP_OneAndHalf – DP method, which allows 1.5-time warping.

shDP_optimalLimitTwoTimeSearch – DP method used in the sliding analysis of the signal during continuous speech recognition. It allows to find the optimal decision at the unknown (floating) length of the word.

shMakeClassicDP_Vectors - DP method, which allows two-time warping. It works with unquantized signals.

vFoundOptimalDPVectPath – realizes the search of the optimal path in obtained DP-matrix.

The methods for the gradient descent are:

shGradient_UnBound - admits unlimited deformation of speech temp.

shGradient_TwoTime – admits two-time deformation.

"**WordCreator**" – realizes the parametric representation of speech signal in the format, which is required for further processes of recognition/understanding and also the saving information in the databases. The class contains instance of the classes "FeatureCoding" and "VectQuant". The main methods of the class are:

vInitializeAcousticProcessor – initialization of the methods for parametric representation of speech signal;

shGetWordFromSignal –transformation of the speech signal into the chain of the numbers of clusters using methods of the feature extraction and clusterization.

"**WordProcessor**" – child of the "WordCreator". It provides work with acoustical-lexical databases, contains the methods for initialization, loading, saving databases, and also accomplishes the databases processing (filling, clearing fields).

"**SlideWordQuest**" – provides the continuous speech recognition by the sliding analysis. For work with acoustical-lexical databases it contains the instance of the "WordProcessor" class. The sliding analysis of continuous speech is accomplished in on-line mode or in the batch processing mode (recognition by earlier created databases). The base methods of the class are:

vParameterInit – setting the main parameters of the sliding analysis;

vSetWordProcessor – setting the word vocabulary;

vLoadConfiguration – loading the main parameters of the sliding analysis from the configuration file;

shSlideAnalysis – accomplishes the sliding analysis of input signal. Parameters are:

shSignalLen – length of the signal;

pSignalData – memory address, where the input data are stored.

The result of the recognition is stored in pPhraseHypList structure, which contains the list of the hypotheses with the following information: shSyntList – words chain, chSyntNumber – number of words in the chain, shAcoust – acoustic estimation of the hypothesis.

vBatchProcessing – accomplishes speech recognition in the testing mode. At that the following databases are used: the word vocabulary and list of the entered test phrases. Each phrase is consecutively processed and the list of the most probable phrase hypotheses (words chains) is displayed and saved in the file.

"**WordRecognizer**" – realizes the word recognition in the on-line mode. It contains an instance of the "WordProcessor" class. The class provides recognition of the words recorded separately as well as the continuous speech.

"**TxtVoc**" – provides the work of the lexical database, accomplishes initialization, loading, saving and processing of the word vocabulary.

"**SituatContext**" – child of the "TxtVoc" class, provides the work with databases of high-level information (semantic-syntactic, situational-pragmatic).

"**IntegralProcessing**" – realizes interaction of different kinds of knowledge during understanding of speech. The base methods are:

shContInsightProcessing – method realizes the understanding of the recorded phrase and uses the following additional methods:

ulContAssociateProcessing – associative processing;

ulContPragmaticProcessing – pragmatic processing.

The class contains the set of the structures required for storage and processing of the phrase hypotheses. The result of the speech processing is displayed or saved in the file.

"**Understander**" – child of the "IntegralProcessing". It combines all the methods of classes of speech understanding model (signal recording, parametrical representation, vector quantization, dynamic programming, high-level and integral processing, databases adjustment). Besides this class provides the user interface.

The developed software codes are available well. It has sufficiently detail comment and description for common usage.

5.1.2. Testing and debugging the developed software

Adjustment (debugging) and testing are very important stages of development of software. The adjustment of program and algorithmic errors, which were missed at these stages, is more difficult and more expensive process at stage of software exploitation.

Any experimental system like this demands numerous adjustments of initial data, arising both during development and debugging of a voice control system. At that, the debugging process can be performed in several phases: (1) - autonomous system debugging; (2) - debugging jointly with an emulator of a control object; (3) complex debugging together with a real control object.

Debugging the produced software was made using standard methods of debugging (program tracing, use of check points, messages logs, mechanism of exceptions etc.) as well as specialized software: integrated development environment Microsoft Visual C++ 6.0, NuMega BoundsChecker Visual C++ Edition 6.5 and NuMega TrueTime for MSVC++. Debugging the produced software system was made both in autonomous regime and in jointly with the aircraft emulator.

During last quarters the developed software system was in the stage of alpha-testing in so-called private regime (without involving outside testers). Principles of black box (test data are generated based on functional specification of a program or a module) and white box (based on structure of a program i.e. passage of operators, branches, etc.) [71] were used as main approaches for testing of the algorithms and programs. Moreover, such methods as: boundary conditions testing, testing of the transitions between software regimes, data integrity tracing, compatibility and transformation of data formats, load testing, were used.

During testing the separate modules, main algorithms (sliding analysis of continuous speech, dynamic programming, Fast Fourier Transformation, calculation of spectral-difference features, speech endpoint detection, etc.) and their assemblies were checked.

Moreover, using NuMega TrueTime for MSVC++ the testing of productivity of separate modules and algorithms was made for detection and adjustment of parts of the software, which are especially critical to time of execution.

5.2. The research prototype of the speech understanding model and database manager programs

In this section the user manuals for work with the research prototype of speech understanding model and additional programs for adjustment of the speech databases (situational, associative, lexical and acoustical) are adduced. The process of the database adjustment is provided by the corresponding management programs, which are described more detail in the following sections.

5.2.1. The research prototype of the speech understanding model

In this section the research prototype of the speech understanding model and description of its main elements are presented.

5.2.1.1. Description of the dialogue window

The main dialogue window of the elaborated research prototype of speech understanding model is presented in Figure 44. In the main dialogue window 8 basic units of control and visualization can be mentioned:

- **«Audio Display»**
- **«Audio Setup»**
- **«Controls»**

- **«Situation Context»**
- **«Internet command»**
- **«Hypotheses list»**
- **«Databases»**
- **«Syntagm/Phrase list»**

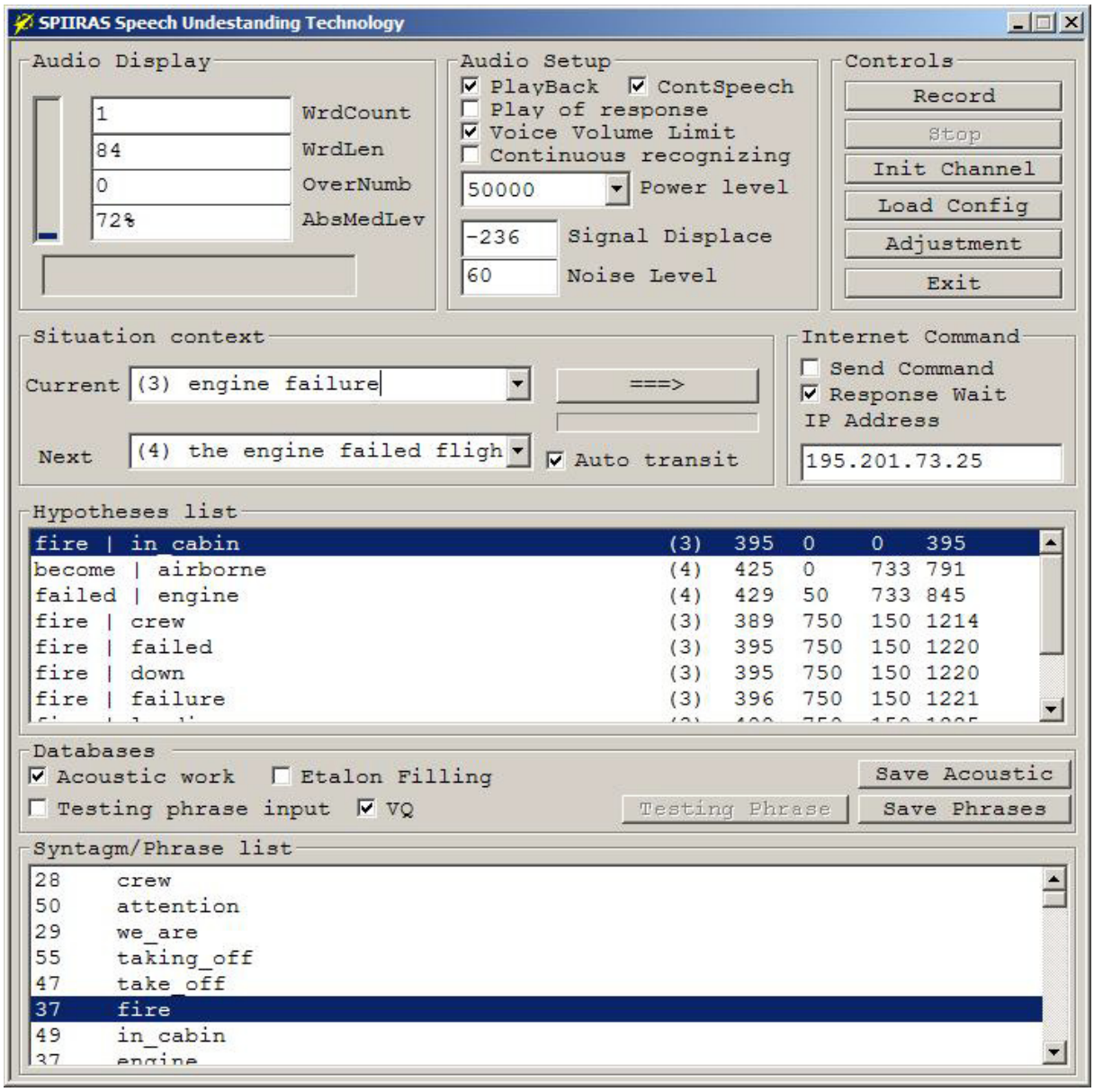


Figure 44. The dialogue window of speech understanding model

Audio Display. The unit provides the visualization of recording process, display of the warning messages and the demonstration of characteristics of recorded signal. The module contains the indicator of volume of signal, the window for the messages about upper limit crossing the voice signal. The recorded signal is characterized by the following parameters:

- «**WrdCount**» – the number of words (syntagmas);
- «**WrdLen**» – the syntagma length in number of segments;

- «**OverNumb**» – the number of samples, value of which exceeds the maximal allowable signal amplitude;
- «**AbsMedLev**» – the average level of the recorded signal.

Audio Setup. The unit provides the adjustment of parameters of the microphone channel and has some options for speech playback. Using the check boxes a user can turn on or turn off the following options:

- «**PlayBack**» – allows to playback the recorded signal;
- «**Play of response**» – allows to playback the wav-file corresponding to the understanding result;
- «**Voice Volume Limit**» – allows to check the signal level displaying the messages «**Speak softer!**», «**Speak louder!**»;
- «**Continuous recognizing**» – allows to perform a continuous mode of control or testing.
- "**ContSpeech**" –allows to choose the mode of the speech input (continuous speech input or isolated word input).

For optimal detection of the speech boundaries the following parameters are used in the model:

- «**Power Level**» – the threshold value of speech segment energy;
- «**Signal Displace**» – the signal displacement relatively to zero. This parameter is automatically adjusted;
- «**Noise Level**» – the noise level. It is automatically adjusted.

Controls. The unit provides the management of recording, recognition, understanding and testing processes. The unit contains 5 buttons for control of the following processes:

- «**Record**» button starts signal recording. In all modes the signal recording and speech detection algorithm are started;
- «**Stop**» button stops any processes (recording mode, finish of phrase recording, demonstration of syntagm recognition results, etc);
- «**InitChannel**» button initializes the recording audio channel for adjustment of «**Signal Displace**» and «**Noise Level**» parameters;
- «**Load Config**» button loads the model parameters from the «**SR_Setup.cfg**» configuration file;
- **«Adjustment»** button activates the process of model adjustment (speech databases) to the new applied task. The keystroke displays the additional dialogue window, in which a user can choose the required mode of the adjustment: integral adaptation of all the databases or adjustment of the concrete database.
- «**Exit**» button completes the work with the understanding model and closes the program.

Situation Context. The unit shows the transitions according to the situational diagram and contains the following elements:

- «**Current**» element shows the title of the current situation. But it is possible to select and reset the current situation from all possible situations using this element.
- «**Next**» element shows the title of the next situation. It is appeared as a result of the understanding. Also the manual selection is possible from the list of all hypotheses, which are obtained by the understanding process.
- «➔» - this button accomplishes the transition of the understanding model from the current situation to the next situation.
- «**Auto Transit**» check box allows to accomplish automatic transition (without result confirmation) to the next situation, which is obtained as the best understanding hypothesis.

Internet command. The unit provides the control of a remote object (a robot, an aircraft model, a machine) by Internet. The unit contains the following elements:

- «**Send Command**» check box turns on or turns off the unit;
- «**Response Wait**» check box allows to set the understanding model in the mode of waiting the response command from the operated object. The option synchronizes the following processes: input of the voice command and execution of the command in the operated object.
- «**IPAddress**» displays the IP address of PC, which controls the operated object. is set in this field.

Hypotheses list. The unit demonstrates the results of recognition and understanding processes. The best hypotheses list with the corresponding estimations is displayed.

Databases. The unit is used during the creation and correction of acoustic databases of syntagmas and phrases (files with *.acs extension). This unit contains the following elements:

- «**Acoustic Work**» check box provides filling the acoustic templates of syntagmas. In the «turn on» position the recording of syntagmas templates is accomplished. The recording process is automatically stopped using speech endpoint detection algorithm.
- «**Etalon Filling**» check box allows input of the syntagmas templates (in the «turn on» position). In the «turn off» position the acoustic recognition of the recorded syntagm and the phrase understanding are realized.
- «**Save Acoustic**» button saves opened acoustic database of syntagmas into the file.
- «**Testing Phrase Input**» check box provides input of the testing phrases and testing of the understanding algorithm in order to optimize the model parameters.

- «**Testing Phrase**» button performs the mode of testing the understanding algorithm.
- «**Save Phrases**» button saves the current acoustic database of phrases.

Syntagm/Phrase list. Depending on the options this unit demonstrates a textual list of the syntagmas or phrases, or transition into the terminal situation obtained as a result of the understanding process.

5.2.1.2. Adjustment of the audio channel

The quality of the understanding model strongly depends on skills of a user to operate with microphone and correct settings of the audio channel. For optimal work of the model it is required to adjust the audio channel. It allows to detect the syntagma boundaries more exactly and avoid some mistakes at the understanding stage. "**Windows Volume Control"** tools or the model settings can be used for this aim.

Parameters of the Windows Volume Control. Depending on the microphone and acoustic environment the amplification level «**Record Control/Microphone**» (included additional amplification «**Mic Boost +20 dB»**) must be set in the optimal value. The amplitude of signal must be used sufficiently fully but without exceeding the upper boundary.

Adjustment of the model parameters. The model has own setting of the audio channel. The adjustment of the parameters is automatically accomplished during the model initialization or using «**Init Channel**» button. As a result of the adjustment the following parameters are optimized: «**Signal Displace**» (the signal displacement relatively to zero) and «**Noise Level**». If the noise level is too much then it is necessary to increase «**Power Level**» value and speak louder.

5.2.1.3. Word recognition mode

The following options must be set in the specific position for work in this mode:

- «**Acoustic Work**» - «turn on» position;
- «**Etalon Filling**» - «turn on» position.

Then in the «**Syntagm/Phrase list**» list it is necessary to set the cursor on the specific syntagma and click the «**Record**» button. After recording the first syntagma the process is automatically stopped and the recorded signal is transformed into the acoustic template (description by features). «**Save Acoustic**» button saves the syntagma templates in the file.

Word recognition mode allows to test the acoustic database of syntagmas. The following options must be set in the specific position for work in this mode:

- «**Acoustic Work**» - «turn off» position;
- «**Etalon Filling**» - «turn off» position.

Then it is necessary to click the «**Record**» button and input a syntagma using microphone. The result of recording is displayed in the «**Audio Display**» unit and the results of syntagma

recognition (list of the best hypotheses with the corresponding estimations) are displayed in the «**Hypotheses list**» unit.

The «**Continuous recognizing**» option can be set in the «turn on» position for the continuous testing of the syntagma recognition. In this case the model automatically transits in the recording mode on completion of the recognition process and demonstration of recognition results.

5.2.1.4. Phrase understanding mode

In this mode a user pronounces a phrase using the microphone and results of the understanding are displayed on the screen and saved in the «**Result.txt**» file.

The following options must be set in the specific position for work in this mode:

- «**Acoustic Work**» - «turn off» position;
- «**Etalon Filling**» - «turn off» position;
- «**Testing Phrase Input**» - «turn off» position.

Furthermore the following options can be used:

- «**Auto Transit**» checkbox allows to accomplish automatically the transition from the current situation to the next situation, which is obtained in the result of understanding.
- «**Continuous recognizing**» checkbox (in the «turn on» position) and if the «**Auto Transit**» checkbox is set in the «turn on» position then the model automatically passes to the recording mode after finishing the understanding process and demonstration of the results.
- "**ContSpeech**" checkbox allows to choose the input mode (continuous or isolated speech).
- «**Send Command**» checkbox in the «turn on» mode the model allows to send the understanding result to the PC with concrete IP address.

To start the model it is necessary to click the «**Record**» button, input a phrase and click the «**Stop**» button.

If the «**Auto Transit**» is checked then the result of acoustic recognition is displayed in the «**Hypotheses list**» unit and the model waits the confirmation of continuation. The confirmation can be made by the «**Stop**» button after that the list of the best phrase hypotheses is displayed in the «**Hypotheses list**» unit. For each hypothesis the following estimations are displayed:

- acoustic estimation of the recorded phrase;
- associative estimation;
- pragmatic estimation;
- integral estimation.

In case of the mistake of the understanding process it is possible to set correct next situation by the «**Next**» element. In order to confirm the understanding result and realize the transition to the next situation it is required to click the «➔» button.

If the «**Auto Transit**» is checked then the demonstration of understanding results and automatic transition into the next situation are automatically performed on completion of the phrase understanding.

If the «**Continuous recognizing**» and «**Auto Transit**» options are checked then on completion of the understanding process the model automatically passes to the recording mode.

Using «**Load Config**» button the model parameters can be updated from the «**SR_Setup.cfg**» configuration file.

Besides the developed research model of the speech understanding can be easily adapted to concrete applied task by integral adaptation. The process of the integral adaptation is realized by programs for management of speech databases, which are described in detail in the following sections.

5.2.2. Situational database management program

The module of the situational database management is the most important in the integral adaptation procedure.

The program provides the creation and modification of the situational databases. This database describes the complete object transitions diagram. It is used for estimation of pragmatic correspondence degree of an input phrase to the current situation. The database contains the list of all possible situations and descriptions of the fragments, which define the possible transitions from the current situation to the following ones. The description of each fragment contains the following data:

- A subset of possible transitions from the current situation to the following ones.
- Commands of the transitions (alternative phrases) to the following situations.
- Semantic weights of words/syntagmas for each paraphrasing.

The creation of new situational database is made as follows:

1) Input of list of all situation.
2) Input of the description for all situations (list of possible transitions to other situations).
3) Input of list of alternative phrases for all possible transitions and the semantic weights of syntagmas, which are contained in the phrases.

The main dialogue window of the situational database management program is presented in Figure 45.

The *File* menu contains some items:

1) *Open DataBase* – choice and loading of the situational database;
2) *Save DataBase* – saving the opened database;
3) *Print DataBase* – output of the information from the database on a printer;
4) *Save Database* (*.xls) – converting the database to MS Excel format in order to obtain the demonstrable representation of the situational database.

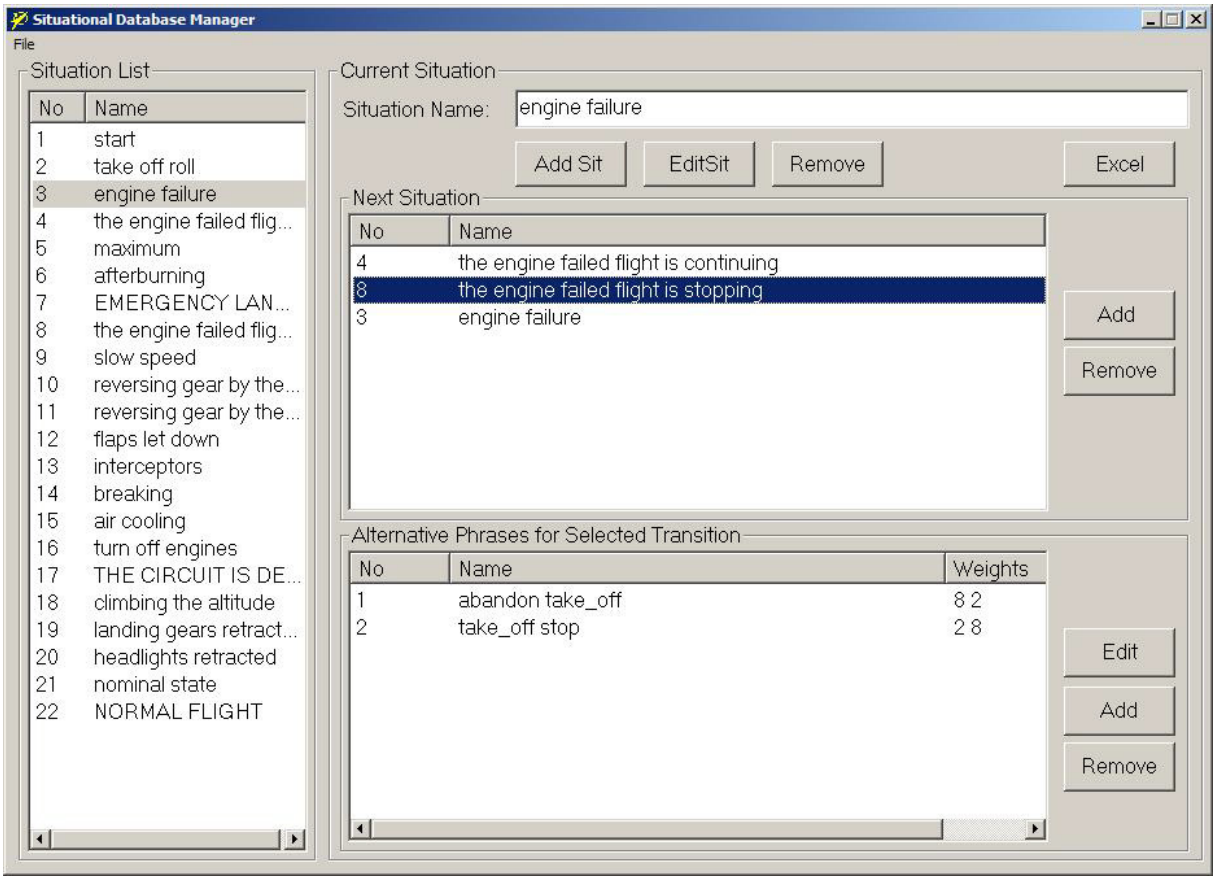


Figure 45. The main dialogue window of the situational database management program

Two main units of control and visualization can be found in the main dialogue window:

- **Situation List;**
- **Current Situation.**

The *Situation List* unit serves for the representation of the list of all situations.

The *Current Situation* unit serves for display of the information about concrete situation (name and number of the current situation). The addition of a new situation and removal of a concrete situation can be performed by the buttons *Add Sit* and *Remove* correspondingly. There are two subunits in this unit:

- *Next Situation;*
- *Alternative Phrases for Selected Transition.*

The *Next Situation* subunit serves for input and visualization of the situations list, into which the operated object can pass from the current situation. The addition and removal of situations can be performed by the buttons *Add* and *Remove* correspondingly.

The *Alternative Phrases for Selected Transition* subunit serves for display of the description of alternative phrases for concrete transition (the phrase partition into syntagmas and the semantic weights of syntagmas in the phrase). The edition, addition and removal of the alternative phrases can be performed by the buttons *Edit, Add* and *Remove* correspondingly.

The main stages of the situational database filling are:

1) **Input of the situations list.** The list of all situations of the operated object must be entered in the *Current Situation* unit. A user must enter the name and number of the situation and confirm the input by *Add Sit* button. After that the situation is displayed in the *Situation List* unit.
2) **Input of the description of all situations.** Each situation is characterized by set of situations, in which an object can transit. In order to input this set the *Select situation to add* dialogue window (Figure 46) must be called by the *Add* button in the *Next Situation* subunit. In this window the possible next situations can be selected. The *Remove* button opens the similar dialogue window for removal of the entered transitions into situations.

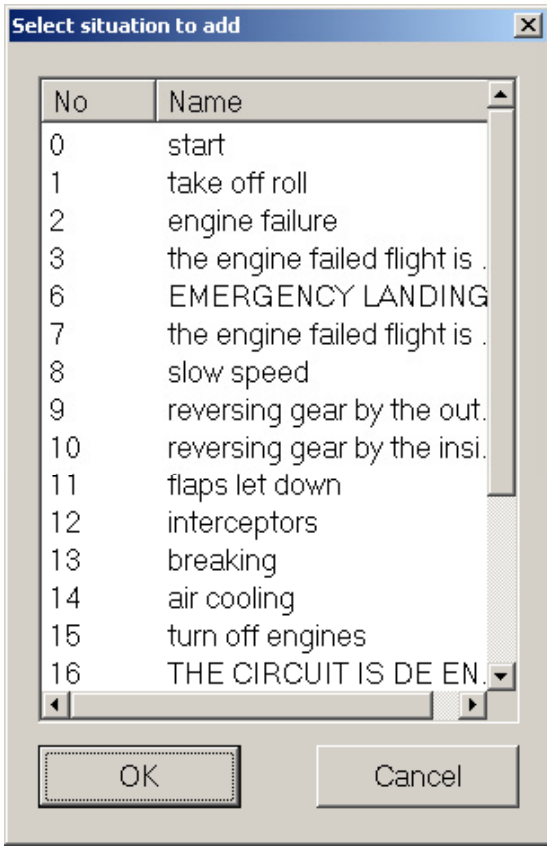


Figure 46. *Select situation to add* dialogue window

3) **Input of the list of alternative phrases.** The transition description is represented as subset of alternative phrases, by means of which this situational transition can be realized. For addition of new phrases the *Edit Phrase* dialogue window (Figure 47) should be called by the *Edit* button in the *Alternative Phrases for Selected Transition* subunit. This window contains two text fields:

1. *Phrase* allows entering the text of an alternative phrase with division into syntagmas. To combine several words in one syntagma the symbol «_» is used;

2. *Weight* allows entering the semantic weights of syntagmas of alternative phrase.

The *Edit* button calls the similar dialogue window *Edit Phrase,* in which the description of the earlier entered phrase can be edited.

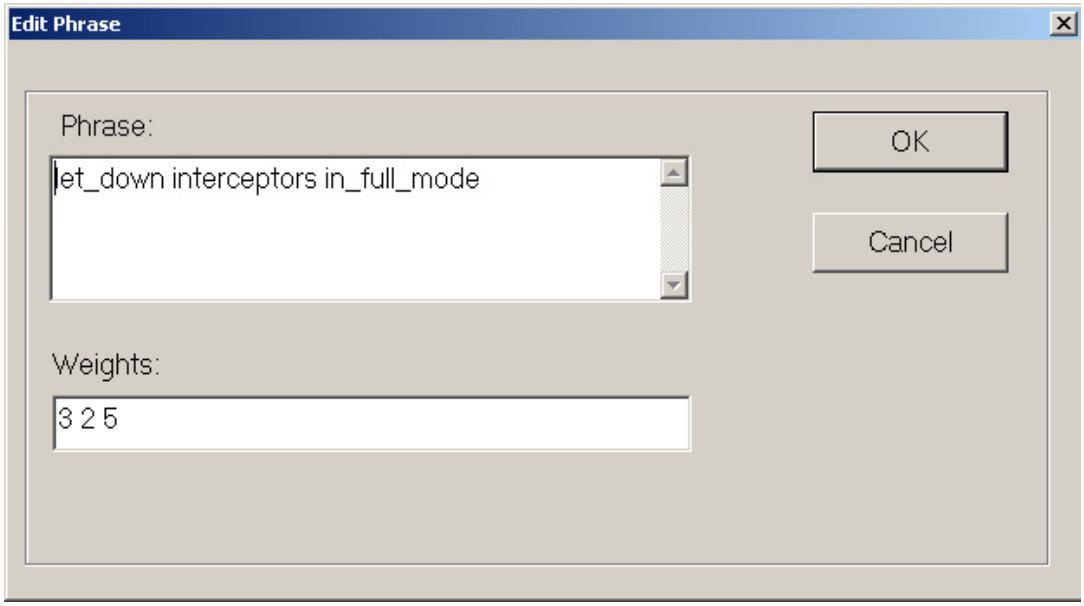


Figure 47. The *Edit Phrase* dialogue window

The items of *File* menu allow to operate with the situational database:

1) to save in binary file;
2) to output on the printer;
3) to convert into MS Excel format.

Moreover during saving the situational database the program fulfills the following actions: (1) creates the syntagma vocabulary; (2) creates the framework of the association matrix using the syntagma vocabulary; (3) creates the simplified version of the association matrix based on the situational database; (4) saves the associative database in the binary file.

5.2.3. Associative database management program

This module serves for expert adjustment of associative estimates, which are contained in the associative database. The associative database contains the compatibility estimations for all the ordered pairs of syntagmas from the syntagma vocabulary. For each pair of syntagmas the associative estimation by 4-score scale is given:

4 – excellent compatibility;

3 – good compatibility;

2 – satisfactory compatibility;

1 – bad compatibility;

Further in the process of speech understanding this estimate is converted into a certain second estimate, which has the nonlinear distribution law chosen by the heuristic way.

The preliminary version of associative database is automatically created, when the situational database is saved. It is realized as follows: (1) the frame of association matrix is created using the syntagma vocabulary; (2) looking through the situational database the following estimates are entered into association matrix: “4” – for discovered pairs of syntagmas, 3” – for the same inverted pairs (i.e. such pairs of syntagmas, which have inverted order), “1” – for other pairs of syntagmas; (3) the associative database is saved as a binary file.

The module of associative database management is the separate program developed using programming language C++ in visual programming environment MS Visual C++ 6.0. Figure 48 shows the main dialogue window of the module, in which the fragment of the association matrix (taken from the developed speech understanding model) is presented.

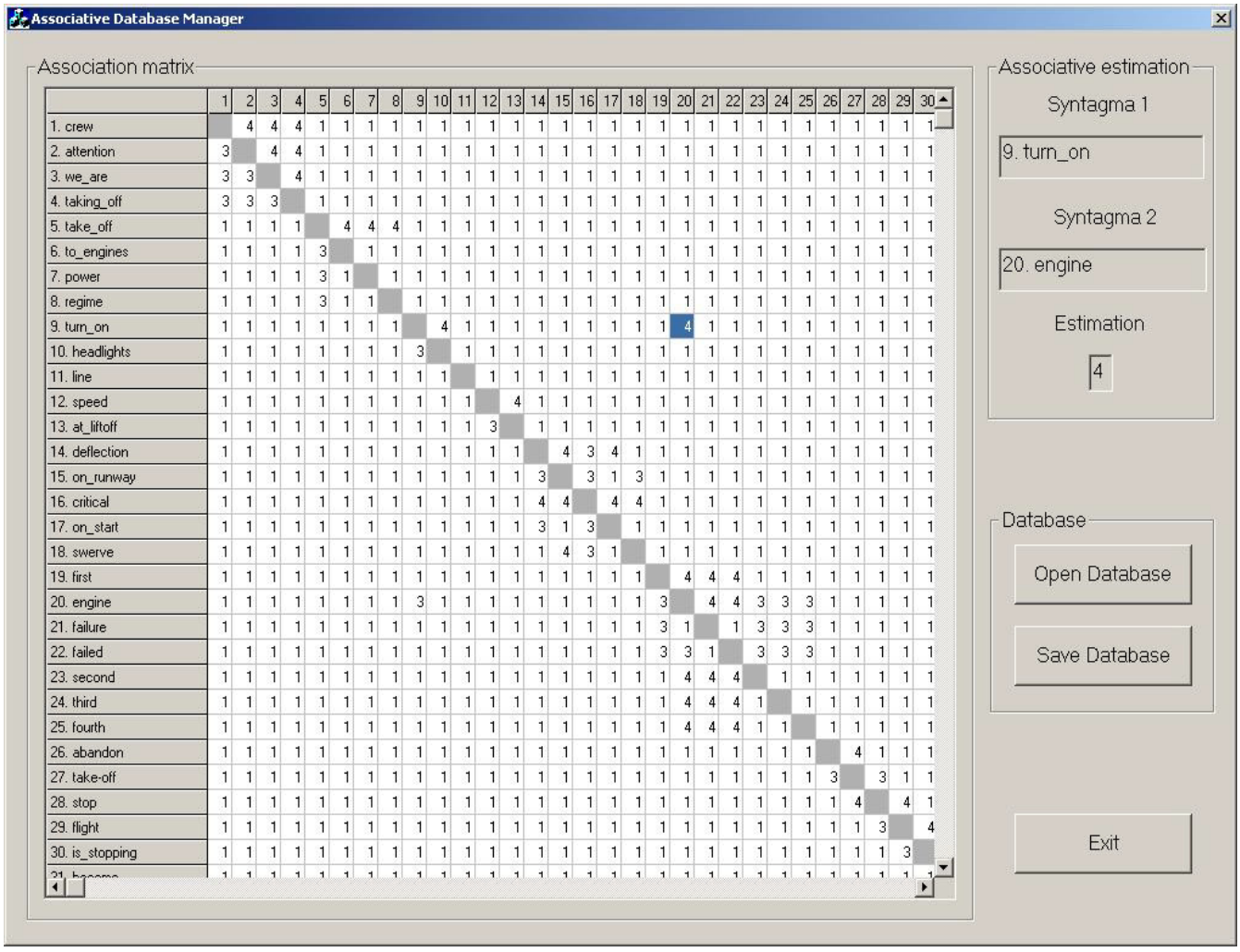


Figure 48. Dialogue window of the associative database manager

In the main dialogue window three units of control and visualization can be found:

- **Association matrix**
- **Associative estimation**
- **Database**

*Association matrix***.** This unit shows in the table form the full matrix of associations between syntagmas presented in the voice control model. In first column of the table the names of syntagmas and their order numbers are displayed. The first row of the table shows the order numbers of syntagmas only. Using keys of cursor control the required row and column of matrix (i.e. corresponding pair of syntagmas) can be selected. On intersection of selected row and column we obtain the value of associative estimate for this pair. The value of associative estimate can be changed by pushing down the numeric key 1, 2, 3, or 4 correspondingly.

*Associative estimation***.** This unit serves for more demonstrable (as compared with full table) representation of associative estimate for selected pair of syntagmas. The fields *Syntagma 1* and *Syntagma 2* display the pair of syntagmas (with their order numbers), selected at current time, and field *Estimation,* which displays the value of associative estimate between these syntagmas.

Database. This unit serves for the operation with files of associative databases. Button *Open Database* allows to choose the file of associative database and represent its contents in the dialogue window. The button *Save Database* allows to save the made corrections in the opened associative database. To have possibility to save the database it is required to set estimates in all the cells of association matrix. After saving the database can be used further in the voice control model.

In order to finish the operation with the associative database manager and close the program it is required to push down *Exit* button.

5.2.4. Acoustic database management program

The module of acoustic database adjustment serves for creation and correction of the database, which contains acoustic templates (parametric description) of all the syntagmas in the voice control model. This module is the separate program developed using programming language C++ in visual programming environment MS Visual C++ 6.0. Figure 49 shows the main dialogue window of the module of acoustic database management program.

In the main dialogue window four units of control and visualization can be found:

- **List of syntagmas**
- **Recording**
- **Loading wav file**
- **Segments**

List of syntagmas. This unit displays the list of names of all syntagmas presented in speech understanding model (i.e. in syntagma vocabulary). The syntagmas, for which the acoustic templates have been already created by a user, are displayed with a numeric prefix, which

means an amount of speech segments occupied by given syntagma (fragment of signal with duration about 11 ms).

The checkbox *Play by selection of syntagma* allows to play back the recorded and saved in WAV file speech signal, which corresponds to selected syntagma. At the same time the corresponding parametric representation of signal is displayed at the unit *Segments.*

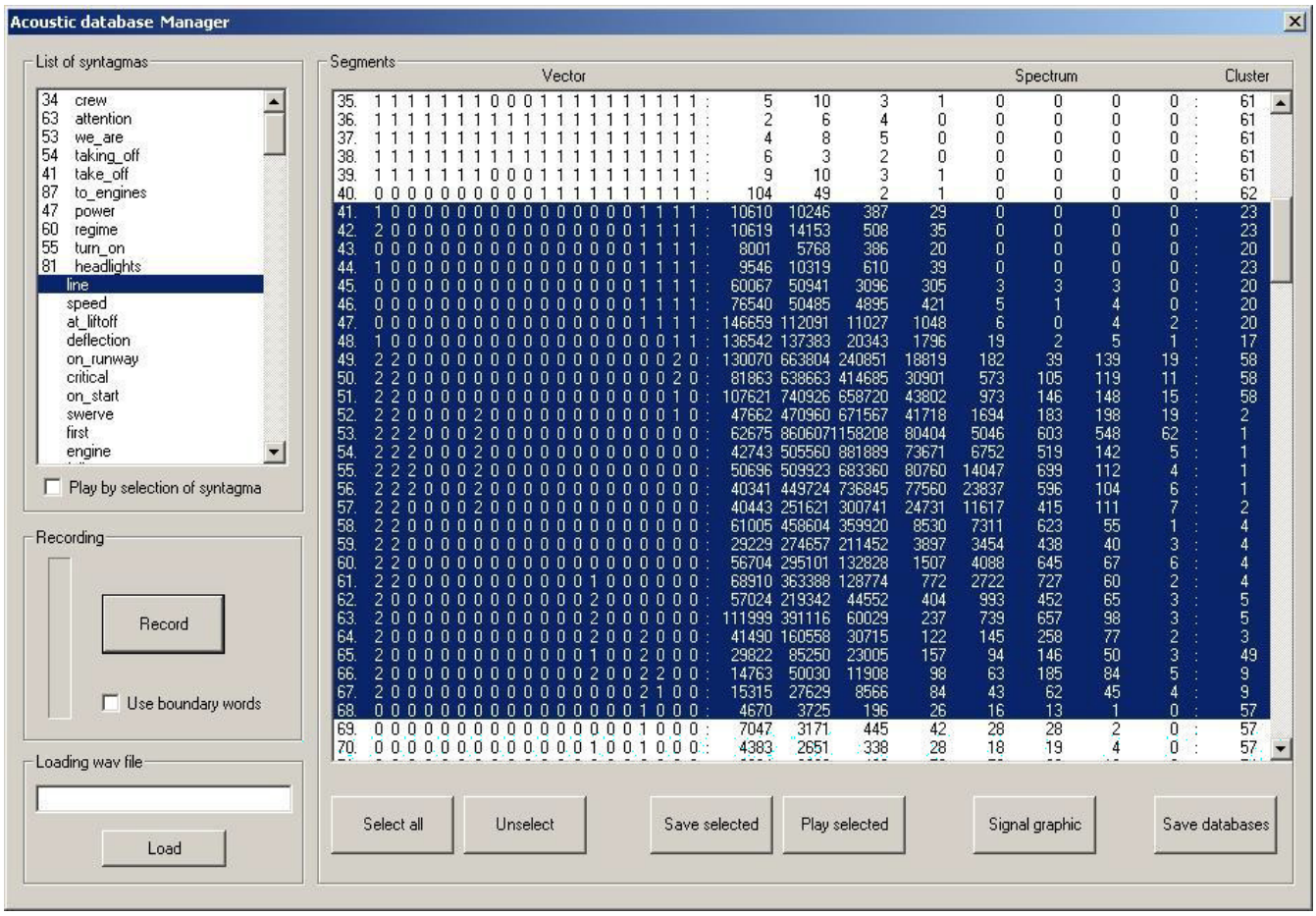


Figure 49. The main dialogue window of the acoustic database manager

Recording. This unit of control and visualization serves for inputting the speech signal through a microphone and consists of three controls:

- The button *Record* initializes and starts the process of recording the signal:
- The checkbox *Use boundary words* turns on the algorithm of speech endpoint detection during the recording. If this checkbox is set on the "turn on" position then the algorithm is used for automatic rejecting the background noise and stopping the process of signal recording. If the checkbox is set on the "turn off" position then the process of recording will be automatically stopped after 3 seconds (this time is enough for recording any syntagma from English or Russian language). In this case the 258 acoustic segments will be recorded from a microphone during 3 seconds.
- The *Progress bar* displays the current level of volume of signal from a microphone. The maximal level of signal corresponds to the value 32000. It should be taken into

account that if during recording the signal level reaches the high boundary of the progress bar (i.e. nonlinear distortions of signal amplitude are present) then it is required to speak softer or decrease the amplification of microphone channel. However if the signal level does not reach the middle of scale then amplification should be increased in order to obtain more qualitative acoustic templates.

Loading wav file. This unit serves for inputting the speech signals saved in WAV files (they must be located in folder "..\Data\Wave"). The *Edit box* inside this unit serves for entering a name of WAV file (without extension). The button *Load* allows to begin the process of loading and playing the speech signal from the file and displaying the parametric representation of the signal in the unit *Segments.* It should be noted too that there is possibility to load WAV file without entering its name in the *Edit box*. It can be made by choosing the concrete syntagma in the*List of syntagmas* (using cursor) and pushing down the button *Load.*

Segments. This unit of control and visualization serves for representation in the table form the parametric representation of inputted speech signal and parametric information manipulation. In the unit *Segments* several controls are located:

- *Table* displays the detail information about each recorded speech segment and contains the following fields:

1. Order number of segment in signal.
2. Vector of spectral-difference (SD) features calculated for this segment. Each vector consists of 20 ternary components {0,1,2}.
3. Value of spectral energy at the outputs of filter bank, which consists of 8 band-pass filters (described above). This information is used for facilitation of the process of manual detection (or adjustment) of speech endpoints.
4. Number of cluster (out of 64 clusters), which corresponds to obtained feature vector for this segment.

Information in this table is displayed on completion of sound signal inputting (from a microphone or a WAV file).

For creation of the acoustic template of syntagma (parametric description) a user have to select the continuous fragment of signal, which corresponds to given syntagma and save this fragment in database. For that it is necessary to select the rows of the table (speech segments). It can be made using both a mouse and a keyboard. By using mouse (left button) it is enough to determine the initial and final segments of signal. In case of usage of keyboard the necessary signal fragment can be selected by the button *Insert*. The button Esc can be used to cancel all made selection.

- The button *Select all* serves for selection of all recorded signal.
- The button *Unselect* serves for unselecting speech segments. It operates analogically as the button *Esc*.
- The button *Save selected* serves for saving the syntagma and corresponding selected signal fragment in temporary memory buffer (out of which the information is written in

acoustic database). After pushing down this button the size of saved syntagma is added before name of syntagma in the *List of syntagmas*. Besides, selected fragment of signal is saved in WAV file (in folder "..\Data\Wave") with the name, which corresponds to the name of selected syntagma.

- The button *Play selected* provides the playback of selected signal fragment.
- The button *Signal graphic* provides the display of inputted acoustic signal. Additional graphical window (analogous as presented in Figure 50) shows the graphics of the signal. Besides this picture can be saved in the corresponding BMP file. The numbers in graphic window show: the minimal value of signal amplitude (negative value), the maximal value of signal amplitude (positive value), the total amount of digital samples in signal. The selected signal segments in the *Table* are displayed in graphics by green background. Thus a user can additionally check the correctness of syntagma endpoint detection.
- The button *Save databases* allows to save the acoustic database with all made changes in the files. The acoustic database consists of the binary file with extension *.acs (acoustic templates using vector quantization) and the binary file with extension *.vec (acoustic templates without vector quantization). The information about each syntagma in *List of syntagmas,* which exists in temporary memory buffers, is written into these files.

To finish the operation with the module and close the program it is required to push down the right button located at window title bar.

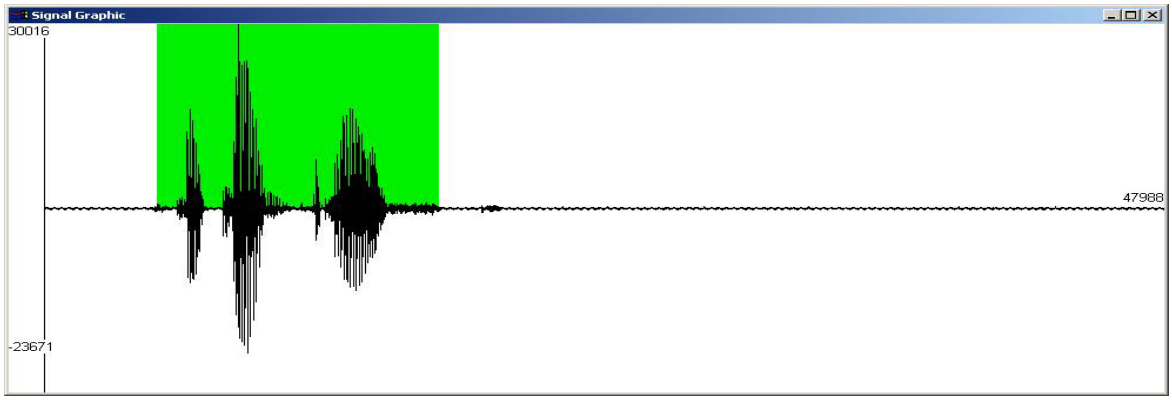


Figure 50. Graphics of the inputted acoustic signal

5.3. Development of the computer emulator of the flying object

It is well known that an emulator is a computer system, which allows modeling the functions and processes taking place in a real control object.

At present emulation is widely used as the important component of the research process of complex objects and it can be also successfully used for the tasks of teaching and training of personnel. The replacement of a complex control object (an aircraft, a satellite, a ship, a

technological process, etc.) by informational or another emulator allows making the significant amount of investigations of an object control system in short time.

During development of an emulator two diverse tendencies collide: (1) to present in an emulator the most of physical, informational and other properties of an object; (2) to present only those properties of an object, which are important for the present stage of investigations. Thus compromise between model complexity and its functional completeness is required here.

The development of the computer emulator of the flying object was performed as follows: (1) compiling the initial data about the control object; (2) adaptation of the situational diagram of the object to the emulation task; (3) selection of the most important parameters for the development of the emulator; (4) development of the emulator and its debugging jointly with the speech understanding module.

5.3.1. Emulation as an important stage of complex objects modeling

The use of simulation models is required in the cases when the use of a real object while designing and debugging the control system is expensive or ineffective. In our work it is required to employ a computer emulator because it is impossible to use the real flying object in the current stage of development. Demands to the emulator, which must be taken into account during development, are:

- Emulator must be a model, which imitates the real object of voice control;
- Emulator must be made as software using the standard means of modeling for PC;
- Emulator must imitate the execution of speech commands;
- Emulator must have obvious visual and sound presentation of the command execution process;
- Emulator must imitate delays of command execution comparable with real ones and allow their variation in certain limits;
- Emulator must have the capability to represent potentially possible emergency situations arising at the object and imitate their elimination;
- Emulator must have the possibility of correcting the functions of the object.

5.3.2. Compiling the initial data for the flying object emulator

The aircraft IL-76 of the Russian civil aviation is the flying object, which was taken as the real prototype of the emulator. The operations for aircraft control were studied by professional methodical manuals [138,151,157] designed for aircrews of different departments, which exploit IL-76. In these books the methods of typical flights by IL-76, the main flight phases and the peculiarities of the flight in extreme meteorological conditions are described. The actions of crew members in emergency situations and troubles during the flight are described too. For additional study of the aircraft functioning we also used the air-simulator Microsoft Flight Simulator 2002.

IL-76 is the four-engined middle transport aircraft. It serves for cargo transportation on Russian and International air-lines of average and long distances. The most important technical characteristics of the aircraft IL-76 are:

- maximum allowed speed in conditions of the normal flight – 600 km/h;
- maximum height of flight – 12000 m;
- maximum take-off weight – 190 tons;
- maximum engine thrust– 12000 kg -s;
- wing area – 300 m^2;
- minimal aircraft crew is 5 persons (aircraft commander, co-pilot, navigator, flight engineer and radio operator).

Below the initial data about the control object required for the development of the aircraft emulator are described in detail.

5.3.3. Representation of the control object as the set of variable parameters

The control object can be represented in the simulation model as set of parameters. The parameters are variables of a certain type, which represent diverse technical components of the aircraft and change their own values by getting respective control commands. The set of parameters values describes the current aircraft state. The process of changing the parameters values shows the flight dynamics exactly and obviously enough.

In our case the state of the real flying object in any stage of the flight can be described in a simplified way by means of the set of parameters presented in Table 21. Each parameter belongs to a certain type of variation (flowing or discrete), has the range of possible values (states), and also has the time interval, which is required for a parameter (real component) to go from one state to another. The time of the value change indicated in the table approximately corresponds to real delays of the execution of flying operations in the aircraft IL-76 (for instance, the retraction of the flaps takes 12-20 seconds). Besides each parameter has a certain initial value (at the start).

Table 21. Main parameters of the flying object

№	Parameter name	Parameter type	Range of values	Time for the parameter change, sec.	Initial value (at start)
1	Flight speed	flowing	0 - 600 km/h	-	0
2	Flight height	flowing	0 - 12000 m	-	0
3	Angle of roll	flowing	-25 - +25	-	0
4	Deflection angle on runway	flowing	0-30 degree	-	0
5	Fuel level	flowing	0 – 80 tons	-	80
6	Deflection from flight course	flowing	0-180 degree	-	0

7	Engine 1 regime	discrete	maximum, takeoff, nominal, low, reverse, failure, off	10	off
8	Engine 2 regime	discrete	maximum, takeoff, nominal, low, reverse, failure, off	10	off
9	Engine 3 regime	discrete	maximum, takeoff, nominal, low, reverse, failure, off	10	off
10	Engine 4 regime	discrete	maximum, takeoff, nominal, low, reverse, failure, off	10	off
11	Headlights state	discrete	on, off	1	on
12	Liftoff speed reached	discrete	yes, no	1	no
13	Contact with ground	discrete	yes, no	1	yes
14	Flaps coordinated	discrete	yes, no	5	yes
15	Landing gear state	discrete	extended, retracted	15	extended
16	Landing gear failure	discrete	yes, no	1	no
17	Flaps angle	discrete	minimal, medium, maximal	15	medium
18	Slats angle	discrete	minimal, medium, maximal	15	maximal
19	Interceptors angle	discrete	minimal, maximal	2	minimal
20	Brake flaps angle	discrete	minimal, maximal	2	minimal
21	Engine 1 generator on	discrete	yes, no	2	yes
22	Engine 2 generator on	discrete	yes, no	2	yes

23	Engine 3 generator on	discrete	yes, no	2	yes
24	Engine 4 generator on	discrete	yes, no	2	yes
25	Engine 1 air bleeding on	discrete	yes, no	4	no
26	Engine 2 air bleeding on	discrete	yes, no	4	no
27	Engine 3 air bleeding on	discrete	yes, no	4	no
28	Engine 4 air bleeding on	discrete	yes, no	4	no
29	Stabilizer failure	discrete	yes, no	1	no
30	Cabin decompressed	discrete	yes, no	5	no
31	Windshield air-cooling on	discrete	yes, no	1	no
32	Distress signal on	discrete	yes, no	1	no
33	Emergency landing	discrete	yes, no	1	no
34	Nose landing gear control on	discrete	yes, no	2	yes
35	Aircraft power on	discrete	yes, no	5	yes
36	Emergency stop	discrete	yes, no	1	no
37	Cabin fan regime	discrete	off, normal, maximal	2	normal
38	Fire signal	discrete	fire, no	1	no
39	Fire in wing	discrete	yes, no	1	no
40	Smoke in cabin	discrete	yes, no	1	no
41	Landing gear air-cooling on	discrete	yes, no	2	no
42	Aircraft balance	discrete	fore, no, rear	5	no
43	Pressure equalized	discrete	yes, no	5	yes
44	Flight regime	discrete	start, normal, stop, continue, landing, braking	1	start
45	“Slow speed” regime	discrete	yes, no	2	no
46	Afterburning regime on	discrete	yes, no	3	no

5.3.4. Situational diagram of the work logic of the control object

Developing the model we are based on the idea of situational control [154], which assumes that the essence of the control process consists in the transition of the object from one situation to a certain following situation. This transition runs under the influence of a human, which possesses all required information for decision-making.

As a result of studying the manual control process by special literature the detailed situational state diagram (Figures 51, 52) had been developed. It represents the flight process at the phases of take-off and climbing the altitude.

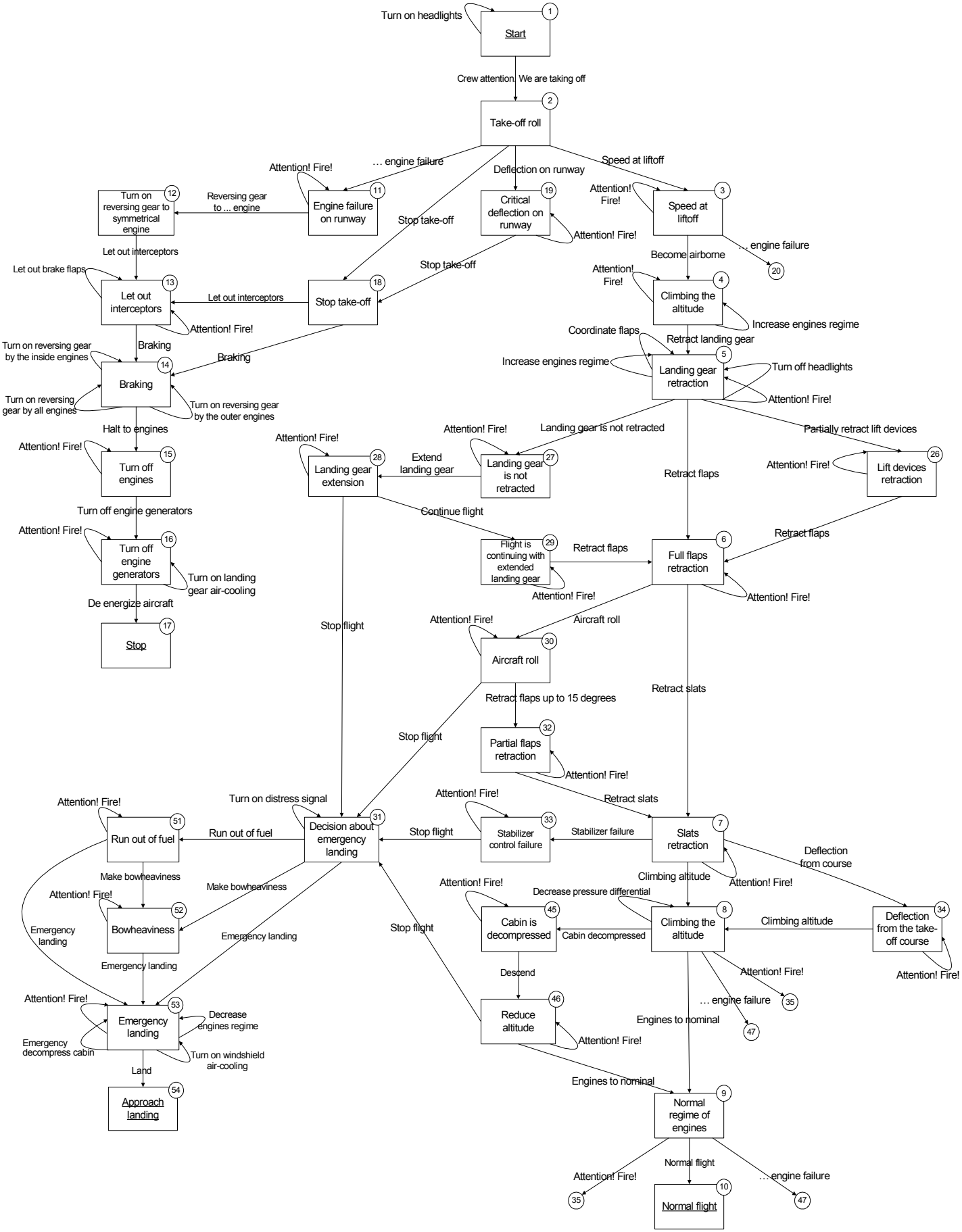


Figure 51. Situational diagram

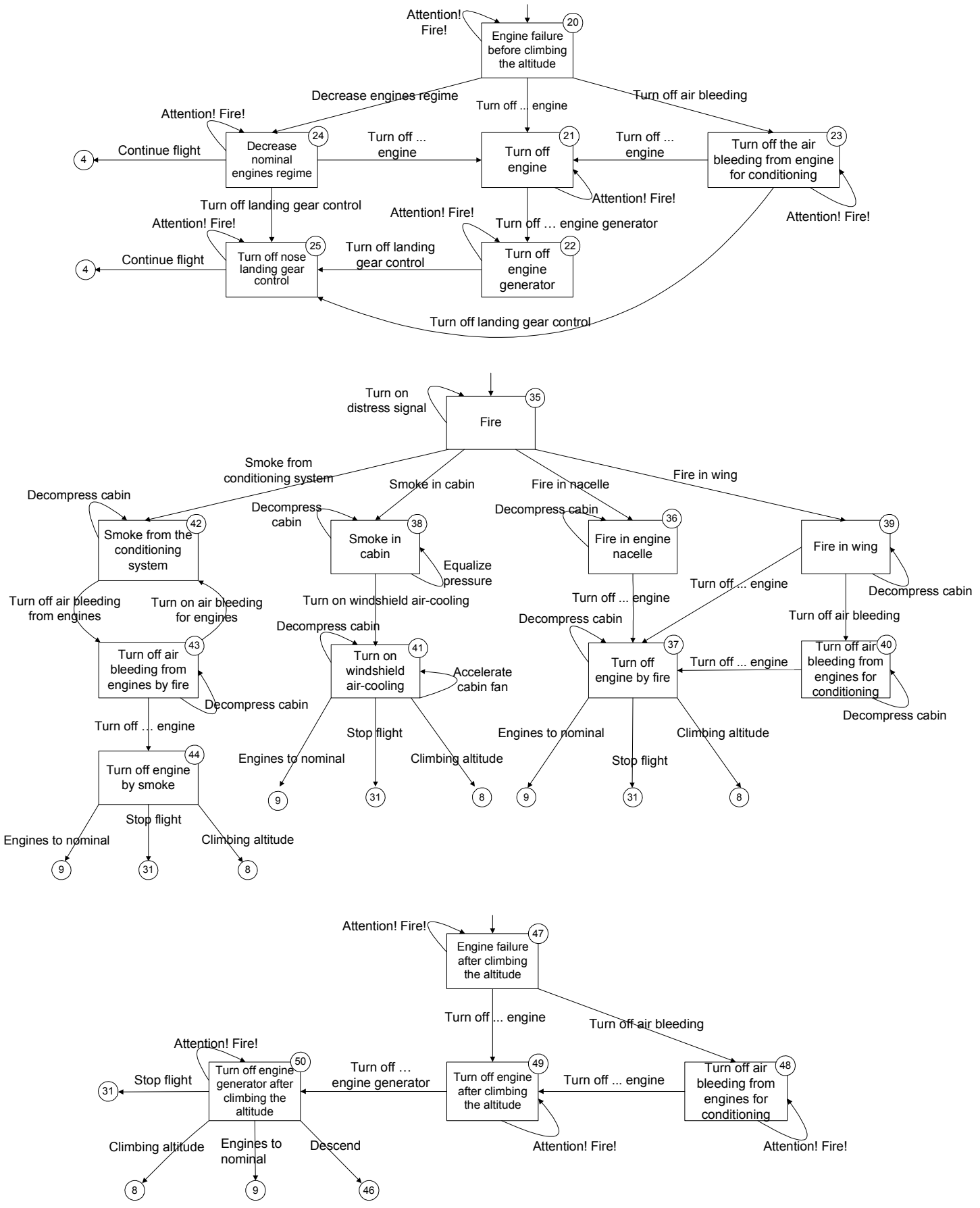


Figure 52. Situational diagram (continuation)

The state diagram shows possible situations for the real aircraft, possible transitions to other situations and also the spoken commands required for the initialization of transitions. Particular attention was paid to investigation the emergency situations such as fire, engine failure, decompression of the aircraft cabin, etc.

For the developed diagram the amount of situations equals 54 (including 3 final); the average coefficient of diagram branching is 4.02; the maximal coefficient of branching is 8; the total amount of different rephrasings is 163; the size of the vocabulary of syntagmas is 152.

The speech databases for speech understanding model have been created taking into account this situational diagram.

5.4. Demonstration model of voice operated flying object

DEMO-version of the speech understanding model provides the demonstration of speech information processing during voice control. A user gives control commands by voice in accordance with the current situation and technical condition of aircraft equipment. In this situational fragment a user can control the aircraft from the take-off phase till the normal flight phase. Also different non-standard situations are demonstrated, which arise during the flight, for instance, “fire in cabin”, “engine failure”, etc.

DEMO-version contains the textual part and the functional one. At the beginning the textual pages are shown, which present main conceptions of the model. Then the functional part is started: a preparation of the model to work, a user acquaintance to the procedure of a speech input and further imitation of the aircraft operation by voice.

The software of the model has been created using visual programming environment Microsoft Visual C++ 6.0 on the programming language C++. Below the minimal requirements to the system configuration (hardware and software) are presented, carrying out of these instructions is required for normal model’s work:

- Personal computer compatible with processor INTEL PENTIUM 200 or higher;
- Microsoft Windows 98 or higher;
- 4 MB RAM or more;
- 5 MB free space on hard disk;

To run the model the following batch files can be used:

- «**Demonstration.bat**» - this file runs DEMO-version presentation and then starts the speech understanding module and the emulator.
- «**Understanding.bat**» - this file runs the complex without textual presentation.

Two main modules (the speech understanding module and the aircraft emulator) and their interaction are described below.

5.4.1. Speech understanding module

For debugging, testing and demonstration of the speech control process of the flying object emulator the research prototype of speech understanding module has been modified. The dialogue window of the modified speech understanding module is presented in Figure 53.

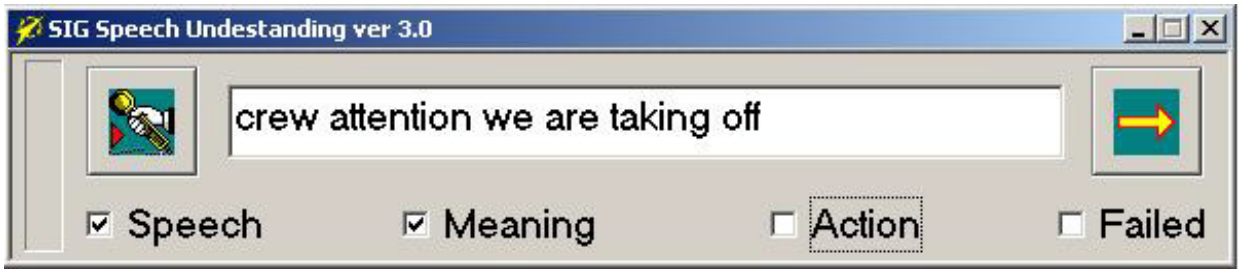


Figure 53. The dialogue window of the modified speech understanding module

This modification includes: (1) the simplified dialogue window (for convenience of work of an untrained user and the demonstration process); (2) means for synchronous interaction between the speech understanding module and the emulator.

Several elements of control and visualization are presented in the dialogue window:

- The button for beginning/ending of the recording process of speech command by a microphone (the left button with the microphone picture);
- The button for transmission of the recognized und understood command to the aircraft emulator (the right button with the communication line picture);
- The text field, which shows the text of the recognized type of meaning (command) or current state of the control process in the situational diagram;
- progress bar, which shows in the logarithmic scale the level of the signalt;
- 4 check boxes, which are automatically set on the “turn on” or “turn off” position at different stages of the command execution process:
  - “Speech” is set in the “turn on” position when the signal input is finished;
  - “Meaning” is set in the “turn on” position when the process of the speech command understanding is finished;
  - “Action” is set in the “turn on” position after getting from the emulator the acknowledgement about the successful command execution;
  - “Failed” is set in the “turn on” position in case of getting from the emulator the message about the failure (impossibility) of command execution.

In the beginning of the recording all the check boxes are set in the “turn off” position.

5.4.2. Aircraft emulator module

It is clearly that control commands are divided into flowing (not discrete) commands (for instance, wheel control, throttle control, etc.) and discrete commands (for instance, landing gear retraction, headlights switching, etc.). It is difficult to give flowing commands orally therefore

the speech input is used for giving discrete commands only. Thus the all flowing operating controls remain under the manual control of a pilot.

Table 22 shows the process of execution of the main spoken commands in the developed aircraft emulator.

Table 22. Visualization of the main speech commands

№	Text of command	Visualization of the command execution
1	- Crew attention. We are taking off - Take off	All parameters “Engine regime” change the values from “off” to “takeoff”
2	- Fire in cabin	Parameter “Fire signal” changes the value to “fire”
3	- Engine failure - Engine failed	Parameter “Engine 1 regime” changes the value to “failure”
4	- Become airborne	Parameter “Contact with ground” changes the value to “no”
5	- Flight is continuing - Taking off	Parameter “Flight regime” changes the value to “continue”
6	- Abandon take off - Take off stop	Parameter “ Flight regime” changes the value to “stop”
7	- Engines to maximum - Take off regime - Give take off	All parameters “Engine regime” change the values to “maximum”
8	- Turn on afterburning - Give afterburning - Afterburning	Parameter “Afterburning regime on” changes the value to “yes”
9	- Turn on emergency landing - Emergency landing - Crash landing	Parameter “Flight regime” changes the value to “emergency landing”
10	- Control wheel to zero - Slow speed	Parameter “slow speed regime” changes the value to “yes”
11	- Turn on reversing gear by the outer engine - Turn on reversing gear by the last engine - Turn on reversing gear by the first and fourth engines	Parameters “Engine regime” for engines 1 and 4 change the values to “reverse”
12	- Turn on reversing gear by the inside engine - Turn on reversing gear by the second and third engines	Parameters “Engine regime” for engines 2 and 3 change the values to “reverse”
13	- Let down flaps in full mode - Let flaps down - Flaps	Parameter “Flaps angle” changes the value to “maximal”
14	- Let down interceptors in full mode - Interceptors full mode - Interceptors	Parameter “Interceptors angle” changes the value to “maximal”
15	- Emergency braking - Emergency - Braking	Parameter “Flight regime” changes the value to “braking”
16	- Turn on landing gears air cooling	Parameter “Landing gear state” changes

	- Turn on landing gears fans - Turn on landing gears coolness	the value to “extended + coolness”
17	- Turn off engines - Halt to engines	All parameters “Engine regime” change the values to “off”
18	- De energize aircraft - De energize circuit	Parameter “Flight regime” changes the value to “stop”
19	- Retract landing gears - Retract wheels - Retract feet	Parameter “Landing gear state” changes the value to “retracted”
20	- Retract headlights - Headlights	Parameter “Headlights state” changes the value to “off”
21	- Nominal - Engines to nominal	All parameters “Engine regime” change the values to “nominal”
22	- Retract flaps - Flaps to zero - Flaps to ten	Parameter “Flaps angle” changes the value to “minimal”

Figure 54 shows the dialogue window of the aircraft emulator AE1. The information about the current state of the flying object (“Aircraft state” group box) and the information about the state of the execution of the last control command (“Command state” group box) are displayed in the dialogue window.

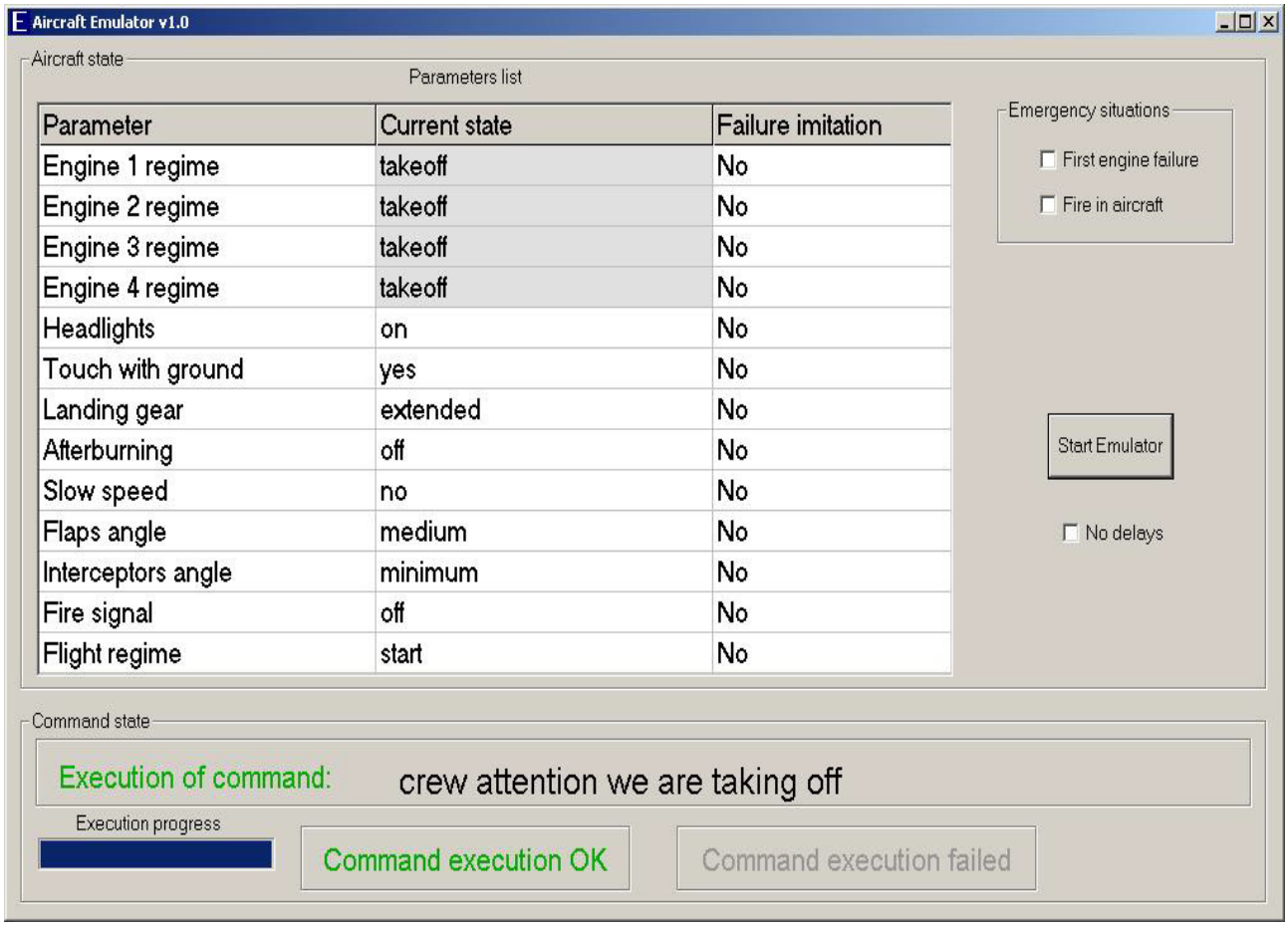


Figure 54. The dialogue window of the aircraft emulator AE1

In “Aircraft state” group box some elements of control and visualization is placed:

- “Parameters list” table of the aircraft parameters with their current states and the information about the presence of failure imitation. Tabular representation of the imitation of the flight dynamics had been chosen as the simplest for a user, untrained in the piloting area;
- “Emergency situations” group box for choice of the emergency situation imitation;
- “Start Emulator” button for initialization of the emulator and transition of the aircraft to the initial state;
- “No delays” check box for turning off the imitation of delays during the execution of commands. When “No delays” is set in the “turn on” position further control commands will be executed without any delays.

The “Command state” group box contains the set of elements for visualization of the command execution process:

- “Execution of command” inscription with the text of the command;
- progress bar, which indicates the time remaining till the end of the command execution;
- “Command execution OK” inscription, displayed in green in case of the successful execution of the command;
- “Command execution failed” inscription, displayed in red in case of the failure of the command execution.

The functioning of the emulator jointly with the speech understanding module is described below.

5.4.3. Interaction between the speech understanding module and the aircraft emulator

The organization of interaction between a real control object and a control system is a complex, but very important task. The solution of this problem influences on the reliability and the quality of work of a control object both in normal functioning regime and during emergency situations. However before integration of the control system into the real control object the testing of the control system with the help of an object emulator must be accomplished in order to eliminate possible errors in a control system and choose the most effective interaction method.

5.4.3.1. The main principles of organization of interaction between the speech understanding module and the flying object emulator

During the flight by the ordinary speech commands are used for the interaction of two persons (at least): an aircraft commander (a main pilot) and a co-pilot (a commander’s assistant). This process is shown in Figure 55. The aircraft commander must first of all (1) know the order of operations with the aircraft equipment to realize the necessary flight plan; (2) transform own intention into the spoken form and pronounce (command) it well. Besides he

must take into account the current situation, external factors, possible emergency situations, etc. The assistant must first of all (1) understand spoken instructions of the aircraft commander quickly and exactly; (2) know all the control elements and the order of operating them. Of course, he must possess the same knowledge as the main pilot to understand the commands as correctly as possible.

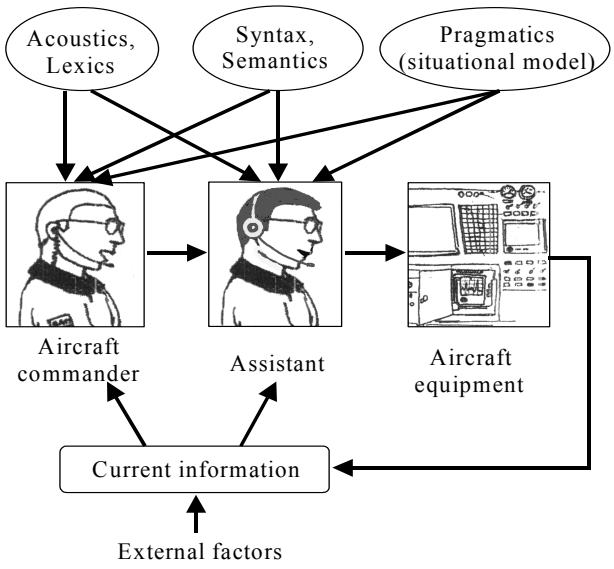


Figure 55. Voice operating with aircraft commander and assistant

Now let us replace the commander's assistant by the automaton, which is able to understand and perform the spoken commands. It is clear, that it must possess the same knowledge as the commander's assistant. It is presented in Figure 56.

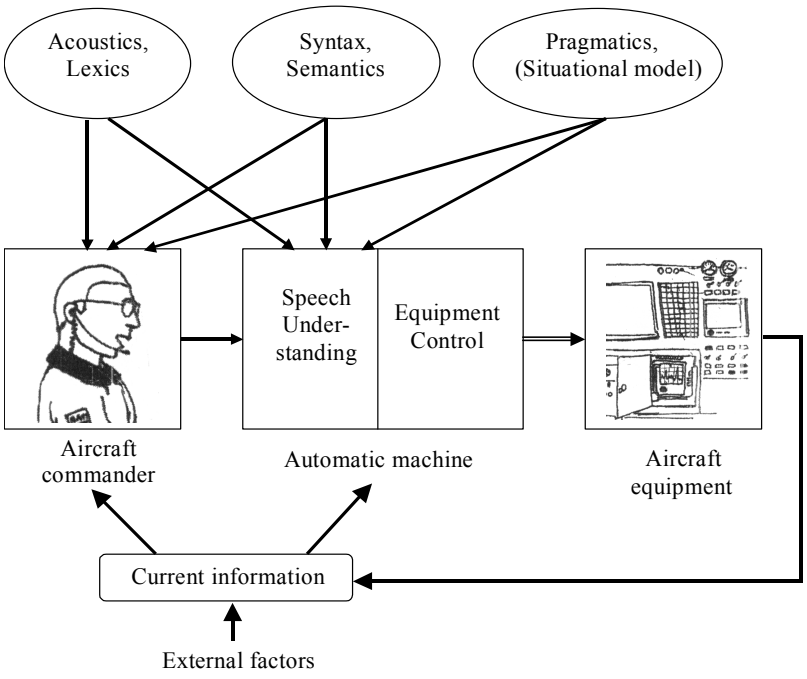


Figure 56. Voice operating with the speech understanding automaton

In case of using the aircraft emulator instead of the real object and speech understanding module instead of automatic machine the Figure 56 is transformed into Figure 57.

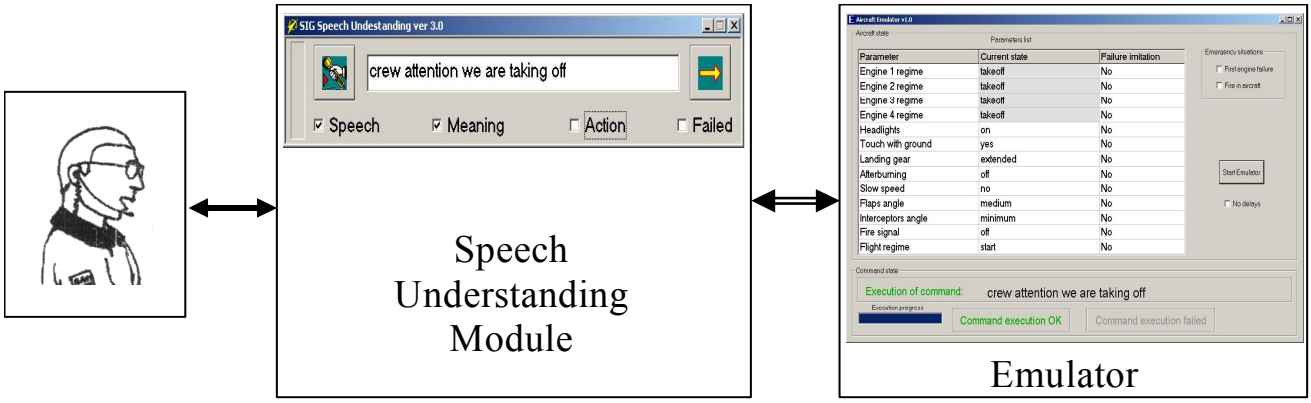


Figure 57. Interaction between the speech understanding module and the aircraft emulator

5.4.3.2. Joint functioning of the speech understanding model and the aircraft emulator

For organization of the data exchange between two programs standard means of the operating system MS Windows were used so called the mechanism of messages queues. The "message" in MS Windows is the notification about a certain event (mouse click, changing window size, pushing a button on the keyboard, etc), which is sent to an application program by Windows or by the other application program. Each of the programs has its own messages queue and the other program sends the new messages into this queue if necessary. In case of receiving any such message the program immediately identifies it and performs the programmed operations.

In the beginning of the work with the developed model a user must run the aircraft emulator. For starting and initializing the emulator it is required to run **AircraftEmulator.exe** executable file and then push the "Start emulator" button in the dialogue window. On completion of the initialization procedure the emulator automatically runs the speech understanding module (**Voice3.exe** executable file**)**, shows list of aircraft parameters on the screen and goes to the regime of command waiting.

After that a user can begin the process of voice control of the aircraft emulator. The sequence of operations of passing the voice command from a user to the object emulator is the following:

1. A user inputs a spoken message (command) into the speech understanding module;
2. The speech message goes through the understanding process and is transformed into the text;
3. The speech understanding module sends (after a user's confirmation) the text of the command to the emulator;

4. The emulator imitates the execution of command during the fixed time;
5. The emulator returns the message, which confirms the successful execution (or impossibility of execution) of the command to the speech understanding module.

To start the recording the speech signal it is required to push the left button (with the microphone picture) in the dialogue window of the speech understanding module. At that the icon with the microphone on the button is changed to the inscription STOP. During the recording of the speech signal the progress bar indicates the current signal level. If the signal is weak (does not reach half of the scale), then for a better recognition quality it is recommended to set a higher amplification level of the signal recording in the "Volume Control" settings of the operating system. To finish the recording it is required to push the left button again. The "Speech" element is set on the "turn on" position and the recorded signal goes to the speech understanding module.

On completion of understanding process the speech understanding module automatically sets the "Meaning" check box on the "turn on" position and shows the text of the recognized type of meaning (command) in the text field of the dialogue window. Besides, the audio file with this command (which was beforehand created by the other announcer) can be heard through the computer dynamic speakers.

To send the command for the execution to the emulator it is necessary to push the right button (with the communication line picture). Simultaneously the message with the text of the command is transmitted to the emulator, the icon on the button is changed to the "STOP" inscription and the speech understanding module goes to the regime of answer waiting.

When the command text is received in the emulator, the command begins to be executed at once. The "Execution of command" inscription is displayed in green and the text of the executable command is displayed. After that if the "No delays" check box is not set on the "turn on" position then the "Execution progress" element begins to indicate the time progress of the command execution. Simultaneously with it the cells with parameters, which must change their values, are selected by gray color. The new values in the cells will be displayed when the "Execution progress" indicator will be set to maximum. If the imitation of the emergency situation was manually inputted into the emulator then some commands cannot be actually executed and the emulator displays the "Command execution failed" inscription in red. If a user did not input the imitation of an emergency situation then the emulator confirms the successful execution of the command by the "Command execution OK" message, which is displayed in green.

After the visualization of the command execution process the emulator returns the answer ("command is successfully executed" or "command execution failed") and goes into the regime of the next command waiting. The speech understanding module sets the "Action" check box in the "turn on" position at the message about successful command execution. Otherwise speech understanding module sets the "Failed" check box on the "turn on" position. In both cases the information about the current state of the control process (in the situational diagram) is

displayed in the text field. In case of successful command execution the transition into the following situation (state) is performed but in the other case the transition is not performed and the speech understanding module remains in the last state.

After the evaluation of the current situation a user (pilot) can begin to input a new spoken command.

Thus in this voice operated flying object a user can control the aircraft emulator at the phases of take-off and climbing altitude by spoken commands, which pass through the process of recognition and understanding in the developed speech understanding module.

6. Perspective research directions in the speech understanding area

This section is devoted to the most vital directions of the research, which are required for the creation of speech technologies for mass use. At this moment the group is conducting the research of speech data organization based on ontologies and in the future we are planning to use the multiagent system for effective collection and updating of the speech and language information. Besides, speech can be used together with other kinds of modality (gestures, handwriting, etc.) for the creation of multimodal interfaces, which will be the base for perspective intellectual systems for human-computer interaction.

6.1. The creation of domain models based on ontology

The situational information is generally used for applied systems oriented at the concrete object area. This information imposes restrictions on the domain and reflects a contextual framework, which is clearly presented in the systems of speech dialogue and voice control by the moving object.

The success of systems, where a user plays the active role in a dialogue, in many respects depends on the following factors: (1) how well the model of domain corresponds to reality; (2) how correctly the system understands an input utterance. These issues are directly connected with the problem of situational information processing.

The development of domain modeling shows that the choice of the data presentation form depends on data specific, internal relations and regularities. The necessity of preliminary analysis of the domain has arisen in order to create the adequate data- and knowledge bases. The attempts to formalize the knowledge about the world and its phenomena in the appropriate for the computer form have led to the appearance of the new investigation area of artificial intelligence – Ontology. Ontology is a universal mechanism of the description of the knowledge structure. It contains a vocabulary, the units of which describe the knowledge and its structure. Besides the vocabulary puts the restrictions on usable semantics. The commonly used definition given by Gruber [31] is «Ontology is a formal explicit specification of a shared conceptualization». In the simplest case ontology is presented as a hierarchy of connected

notions. Moreover the appropriated axioms reflecting the connections between the notions and restricting their interpretation, are used. In Figure 58 the formal model of the ontologies is represented as the system of ontologies of different levels [29,32]:

Top-level ontologies, which include general concepts like time and space, objects, events etc. This type of ontology is domain-independent and should therefore be applicable for all problems and applications.

Domain and Task ontologies, which capture, respectively, a generic domain or a generic task. Either of these types can be constructed by specification of concepts in a top-level ontology.

Application ontologies, which are both domain and task specific. It can be constructed through specification of a set of domain and task ontologies related to the application.

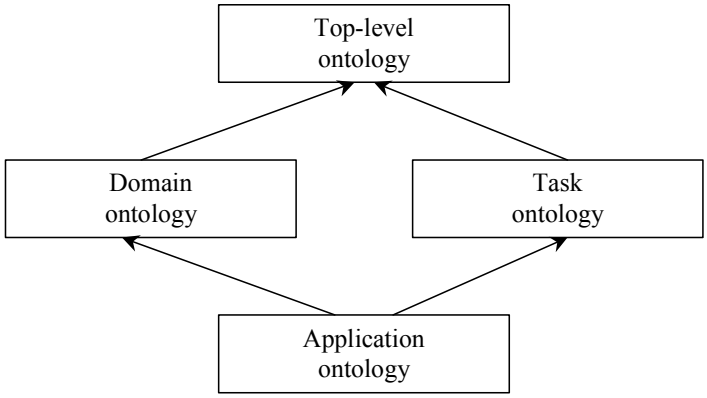


Figure 58. Multi-level scheme of relations between ontologies

In order to construct the domain model semantic and pragmatic kinds of information are used in speech understanding systems. In the developed integral model of speech understanding semantic-syntactical analysis is replaced by associative analysis. Figure 59 shows the process of the creation of databases in the following order: the situational database, the vocabulary of the domain and the matrix of inter-word associations.

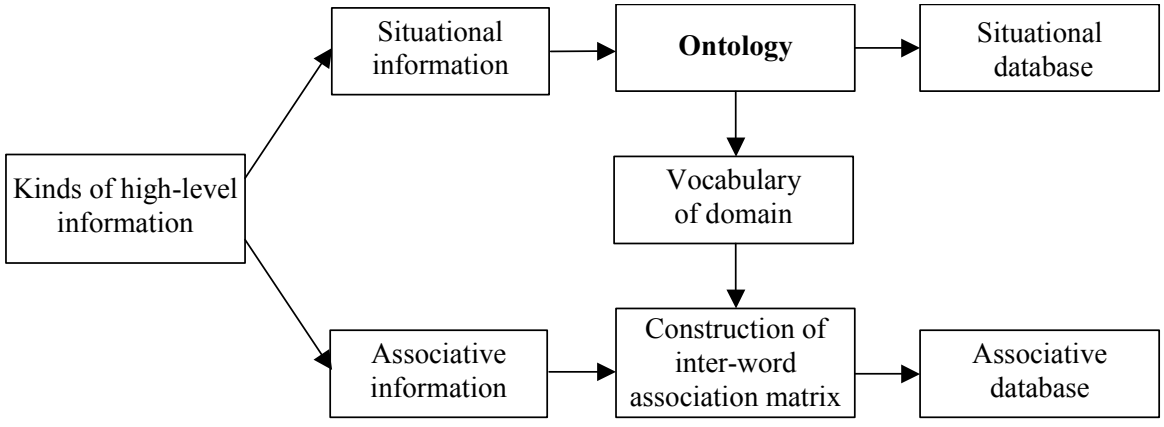


Figure 59. The model of high-level information presentation in the integral model of speech understanding

The creation of high-level databases based on the system of ontologies should start from the analysis of the domain and the development of the top-level ontology. In a voice control system the top-level ontology includes such notions as *situation*, *transition* from situation to situation, the *phrase* initializing this transition, etc. Domain and Task ontologies fill up the high-level framework of the top-level ontology by the situations of moving object control and control commands. During the construction of the ontology for human-computer interaction tasks first of all the problem of high-level information extraction arises. At present in artificial intelligent two base approaches to knowledge extraction are used:

1) The *textual* approach based on expert analysis of natural language (in textual form) from the concrete domain;

2) The c*ommunicative* approach based on the data obtained from experts.

In speech understanding area these approaches have some differences as compared with expert systems. The domain modeling based on natural language analysis is reduced to the construction of the semantic model and oriented to an ideal abstract language. Such model does not take into account variety and polysemy of natural speech.

The means of obtaining knowledge from an expert and his skills play the main role in communicative methods. At that the methodologies of obtaining knowledge and tools for data collection and processing could be borrowed from the expert system area.

Different methods for obtaining knowledge from experts and special texts allow to create the domain model close to reality. However, the problem of adequacy of existent models consists in their disparity to nature of speech dialog. Since the templates of phrases and utterances are formed by text data and grammatical constructions of a natural language.

The appearance of sufficiently reliable recognition systems will lead to the new direction of domain presentation as *modeling of object area based on natural speech data* (Figure 60). In this case forming the domain model will be based on natural speech data.

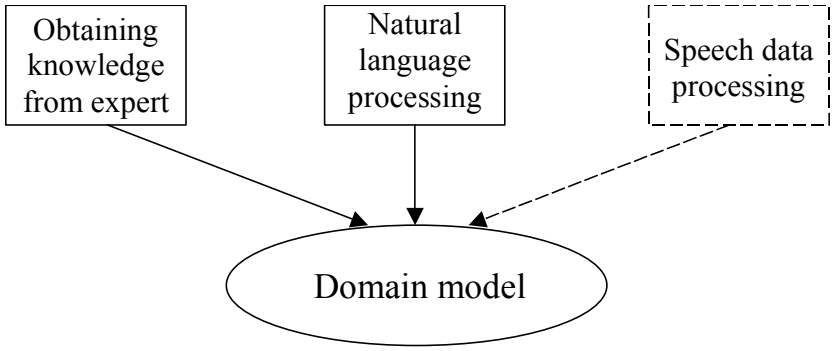


Figure 60. Approaches to creation of domain model

In this project the situational model of a voice operated flying object was created by the expert method. The next step of research is the transition from expert technology to the methods of statistical analysis and quantitative estimation of the data obtained from real speech dialogues. This model will take into account variety and polysemy of natural speech. It will

allow to decide the problem of adequacy of domain and effective interpretation of a speech utterance.

For the construction of the high-level (situational and associational) model in the task of a voice operated flying object using the automated processing of speech data it is necessary to decide the following problems:

1. Choose the kind of control;
2. Describe the model of control;
3. Choose and realize the hardware decision of data collection;
4. Develop the methodology of preliminary text processing;
5. Develop the ontological system for creation of high-level databases.

Each stage requires developmental work and deep knowledge of the concrete domain. The realization of this approach will allow to create the high-level model adequate to speech activity.

6.2. The multiagent system for effective accumulating and updating speech and language data

At present there are many well-known systems for automatic stenography and voice control for usual office programs. However, these systems are used extremely seldom because the first attempts of using such systems have shown that during the dialogue a user does not follow strict grammatical rules and so the speech contains many deviations and inaccuracies. In order to avoid from this situation various software and hardware means were applied. It has increased the technical complexity and the cost of the systems but unfortunately has not decided the problem principally. And a user, as before, refuses to use speech technologies.

So, for creation of the robust speech recognition system the developer must be able to model natural speech. For that the huge corpora of natural speech are required. Besides taking into account that the language is an open system and is permanently changing, we can assume that in the future multiagent systems will be perspective for effective accumulating of the speech and language data. A user will participate in the process of accumulating an updating speech databases simultaneously with work.

The supposed structure of the multiagent system, which can be used for this task, is presented in Figure 61. The devices (a computer, a phone, a laptop, a cellular and other mobiles) capable to recognize the speech have to fulfill the accumulating the speech information obtained during the dialogue between a user and a device. The accumulated information is used in two aspects. Firstly, it is the creation of the local intellect, which allows precise adjustment to the concrete user (adaptation to voice and pronunciation manner). Secondly the information from all users is collected up for the estimation of the current common state of the speech communication and global adjustment of the united speech recognition/understanding system.

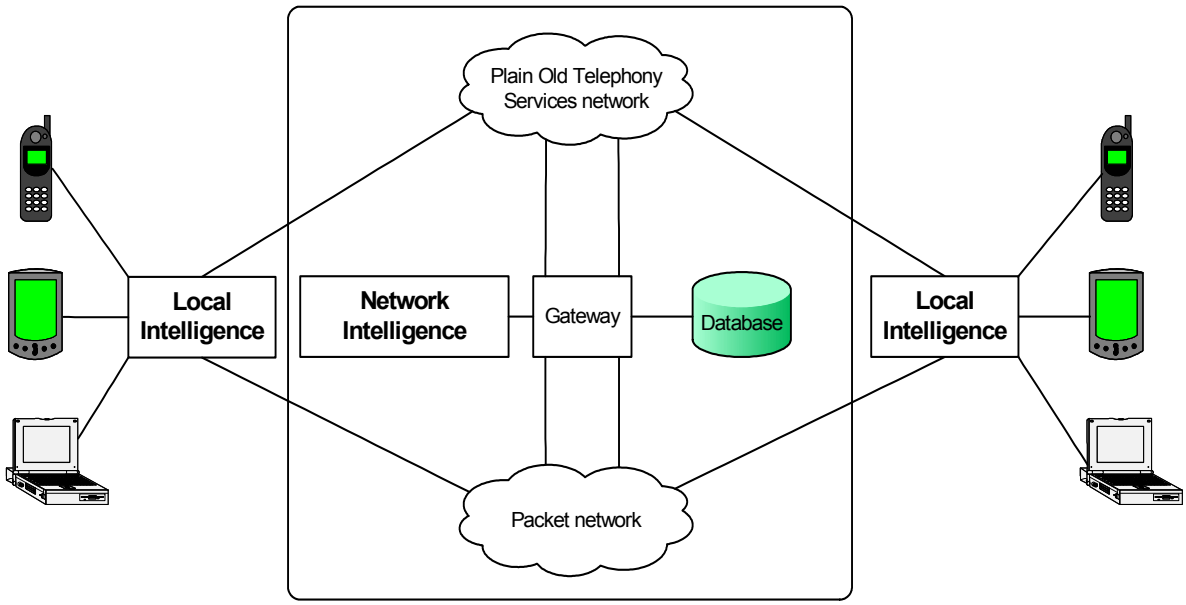


Figure 61. Multiagent system for speech technologies

Such approach will give the required knowledge about the current state of the language and provide high quality of speech recognition/understanding.

6.3. Voice interface in the perspective infotelecommunication systems

Infotelecommunication is a new perspective direction in modern science and techniques, which presents a symbiosis of information technologies and phone industry (telecommunication) [137]. At present the infotelecommunication market in the world is developing very quickly. One of the most perspective directions in this market is the elaboration of new services and systems, which could maximally use various communication abilities of a human and, first of all, in natural speech. Some examples of perspective applications are [21]:

- automation of operator services;
- automation of directory assistance;
- voice dialing;
- reserve directory assistance;
- voice messaging;
- voice response systems;
- extended banking services, etc.

It is necessary to mark that the usage of speech recognition technologies in telecommunication companies gives many advantages:

– reduction of financial expenses of telecommunication companies (by automation and improvement of existent services);
– attraction of new clients (owing to the usage of new intellectual services);

- satisfying the needs and wants of clients (comfort of the information access: in any time, from any place, by simple and natural interface);
- reduction of the expenses of clients (time saving);
- improving the image of a company (in comparison with competitors, which do not use these technologies);

These and other positive factors require the fast development and application of speech technologies in infotelecommunication area.

6.4. Multimodal interfaces for the task of human-computer interaction

Multimodal interface is the kind of natural interface, which processes in a coordinated manner two or more user input modalities such as speech, handwriting, manual gestures, gaze, head and body movements, etc [33,83]. The first multimodal system "Put That There" was developed in 1980. It was the demonstration system, which processed speech in parallel with touch-pad pointing. Modern systems are based on two or more parallel input streams, which are capable of conveying rich semantic information. There are active and passive modalities. Active input modalities are used by human intentionally as an explicit command to a computer system (for instance, using speech or gestures). Passive input modalities refer to natural user behavior or actions and accompany the active modalities (for instance, lip movements always accompany with speech pronunciation). They involve user input that is unobtrusively and passively monitored, without requiring any explicit command to a computer.

6.4.1. Main differences between multimodal interfaces and unimodal interfaces

Communication channels can be tremendously influential in shaping the language transmitted within them. Many linguistic features of multimodal language are qualitatively very different from that of spoken or formal textual language. It can differ in features as basic as brevity, semantic content, syntactic complexity, word order, degree of ambiguity, specification of determiners, etc. In many respects, multimodal language is simpler linguistically than spoken language. In particular, comparisons have revealed that the same user completing the same map-based task communicates fewer words, briefer sentences, and fewer complex spatial descriptions when interacting multimodally, compared with using speech alone. One implication of these findings is that multimodal interface design has the potential to support more robust future systems than a unimodal design approach.

Multimodal interface is natural for human communication process. A user can decide which channel for which information he wants to use. The probability of human to have two-way communication with different input/output channels is important too.

The array of multimodal applications has expanded rapidly, and presently ranges from map-based and virtual reality systems for simulation and training, to medical and web-based transaction systems.

In future the information from a large number of different visual, auditory, and tactile input modalities will be recognized and used in everyday activities. The corresponding systems will track and incorporate the information from multiple sensors on the user's interface and surrounding physical environment in order to support intelligent adaptation to the user, task and usage environment.

6.4.2. The most effective combinations of modalities

At present bimodal interfaces are usually used. Some combinations of modalities bring the same information (for instance, speech and lip movements) so different channels can be used to recognize information correctly with more probability and robustness. In other systems different modalities carry different information, which can be integrated at semantic level only (for instance, information from speech and gestures).

The most popular combinations of modalities are: speech and handwriting input, speech and lip movements, speech and manual gesturing or gaze tracking.

Speech and handwriting input. The main destination of such systems is the input on PC or other systems without using of ordinary keyboard. A distinction must be made between textual input and command input. The goal of textual input is to enter linguistic data, either in the form of speech signal, or as text. Handwriting and speech combination can be used to increase bandwidth and reliability in textual input when user spoke and write same text. In some systems is used others combinations such as speech recognition/handwriting synthesis, when user read and correct by pen text which is entered by him with voice or reverse. So a user has a feedback with the system. For example, a user may read letter and correct it with the pen while reading it. Table 23 shows main methods of using speech/handwriting recognition/synthesis combinations [1].

Table 23. Combinations of input-output of speech and handwriting

	Speech recognition	**Speech synthesis**
Handwriting recognition	The system improves the throughput in bulk text entry	The user gets voice feedback on handwriting recognition results
Handwriting synthesis	The user dictates a synthesized handwritten letter	Multimedia text communication by the system

Speech and manual gestures. The main goal of such systems is to give more simple way to manipulate different objects, pointing to them by pen or hand (special case of gesture input). Such systems have advanced more rapidly in their architectures, and have progressed further toward commercialization of applications. This interface is usually used in mapped based

systems. Sometimes for such multimodal systems to point with gesture is more comfortable than to do this with help of words. In Table 24 some characteristics of such systems are presented [33].

Table 24. Characteristics of existing speech-gesture systems

Name of the system	QuickSet	Human-centric word processor	VR Aircraft Maintenance Training	Portable Voice Assistant
Type and size of gesture vocabulary	Pen input, Multiple gestures, Large vocabulary	Pen input, Deictic selection	3D manual input, Multiple gestures, Small vocabulary	Pen input, Deictic selection
Size of speech vocabulary	Moderate vocabulary, Grammar-based	Large vocabulary, Statistical language processing	Moderate vocabulary, Grammar-based	Small vocabulary, Grammar-based
Possible applications	Wireless handheld, Varied map & VR applications	Desktop computer, Word processing	Virtual reality system, Aircraft maintenance training	Wireless handheld, Catalogue ordering

The typical flow of joint processing of gestures and speech is shown in Figure 62. On two first stages the system gets information from input channels. The recognition and understanding of speech and gesture may be parallel and independent. Then system fuses information taking into account context. The last stage is dialogue management, connection with applications and response planning.

Speech and lip movements. Usually these systems are used to recognize speech with more reliability. This type of bimodal interface is natural for human-to-human communication process. The importance of lip movements grows with noise raising. Historically, multimodal speech and lip movement research has been driven by cognitive science interest in intercessory audio-visual perception and the coordination of speech output with lip and facial movements. Several existing multimodal interfaces currently include adaptive techniques for improving system robustness in noisy environmental contexts (for instance, in factory).

Speech and gaze. Gaze and speech bimodality may be used in Attentive User Interface [10], where gaze is the passive modality and speech is active one. It is known that a human usually looks at the object before he begins to speak about it or manipulate. So gaze system gets information about next object of dialog. So it may be more attentive to user's action.

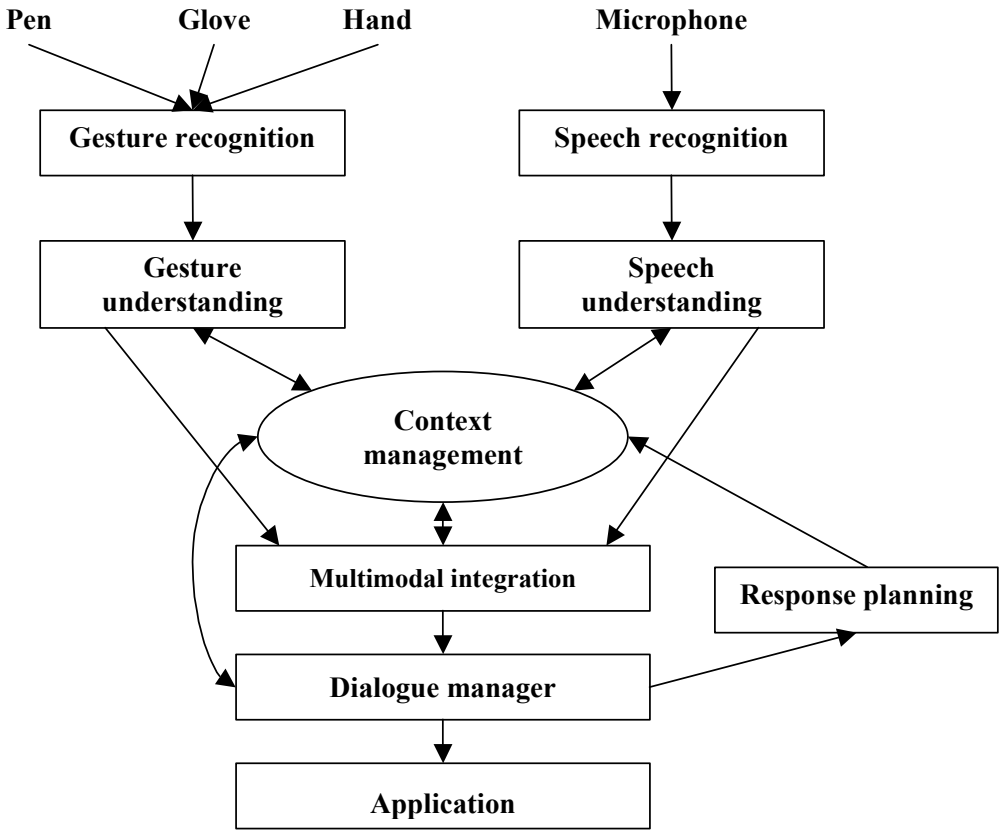


Figure 62. Flow of joint processing of gestures and speech

6.4.3. Perspective directions of usage of multimodal interfaces

Unlike a traditional keyboard and mouse interface or a unimodal recognition-based interface, multimodal interfaces permit flexible use of input modalities. Choice of modality is an important issue in design of multimodal systems. It includes the choice of which modality to use for conveying different types of information, to use combined input modes, or to alternate between modes at any time. Since individual input modalities are well suited in some situations, and less ideal or even inappropriate in others. As systems become more complex and multifunctional, a single modality does not permit all users to interact effectively across all tasks and environments.

The important question, which arises during system development is whether a user communicates unimodally or multimodally. In the case of speech and pen-based multimodal systems, users typically intermix unimodal and multimodal expressions. Predicting whether a user will express a command multimodally depends on the type of action. In particular, users almost always express commands multimodally when describing spatial information about the location, number, size, orientation, or shape of an object. They also interact multimodally when selecting an object from a larger array, for example, when deleting a particular object from the map.

The most perspective systems for using multimodal interface are manipulating systems, robots, or other systems where user can give command by voice, selecting different objects or menu items with pen on table or by hand in natural world.

Also multimodal interface may be useful in telecommunications for development of perspective Personal Digital Assistance, where the usage of ordinary keyboard is difficult. Nowadays systems of speech recognition are used in PDA, their combinations with gestures and handwriting allow to do informational exchange more effectively. There are some research concerning multimodal access to Internet using PDA and WAP [73]. For navigation a user can use gestures and make input of textual information with speech.

So the usage of multimodal interfaces allows to improve the efficiency and competitiveness of diverse applications.

7. Participation in the international conferences. Papers submission to international journals

The results of the research conducted during the project have been presented in the reports and discussed on the following International conferences:

1) Andrey Ronzhin, Yuri Kosarev, Alexey Karpov, Izolda Lee. Elaboration of the intellectual speech interface provided accuracy, robustness and adaptability. International Workshop SPECOM'2003, Moscow, Russia, October 2003, pp. 231-236.
2) Yuri Kosarev, Andrey Ronzhin, Alexey Karpov, Izolda Lee. Approaches to creation of situational databases for integral speech understanding models. International Workshop SPECOM'2003, Moscow, Russia, October 2003, pp. 114-118.
3) Andrey Ronzhin, Yuri Kosarev, Alexey Karpov, Izolda Lee. Recognition and understanding of the continuous speech based on sliding analysis and integral structure of processing. In Proceedings of the Fourth International Scientific and Technical Conference "Intellectual and Multiprocessing Systems IMS'2003", Divnomorskoe, Russia, September 2003, vol. 1, pp. 181-184.
4) Alexey Karpov, Yuri Kosarev, Andrey Ronzhin, Izolda Lee. Development of acoustical features of speech signal robust to variations of signal scale and spectrum deformations, the 3-th All-Russian Conference "Theory and Practice of speech investigations" ARSO-2003, Moscow, Russia, September 2003, pp. 83-88.
5) Yuri Kosarev, Andrey Ronzhin, Izolda Lee and Alexey Karpov Continuous Speech Recognition without Use of High-Level Information. 15-th International Congress of Phonetic Sciences, Barcelona, Spain, August 2003, pp. 1373-1376.
6) A.L. Ronzhin, Yu.A. Kosarev, A.A. Karpov, I.V. Lee. Elaboration of the intellectual speech interface provided accuracy, robustness and adaptability. Samsung Young Scientist Day, Saint-Petersburg, April, 2003. Second award certificate has been received.

7) Yuri Kosarev, Izolda Lee, Andrey Ronzhin, Alexey Karpov Robust Speech Understanding Methods for New Intellectual Applications, AVIOS 2003 Conference, San Jose, California, USA, 2003.
8) Yu.A. Kosarev, I.V. Lee, A.L. Ronzhin, E.A. Skidanov, J. Savage. Survey of the approaches to speech and text understanding. SPIIRAS Proceedings, Issue 1, Vol. 2. – St.Petersburg, SPIIRAS, 2002, pp. 157-195.
9) Yuri Kosarev. Some aspects of Robust Speech Understanding. Invited lecture for the International Workshop SPECOM'2002, St. Petersburg, 2002, pp. 3-8.
10) A. Ronzhin, Yu. Kosarev, I. Lee, A. Karpov. Continuous Speech Recognition Suitable for Robust Speech Understanding. International Workshop SPECOM'2002, St. Petersburg: "Evropeiski Dom", 2002, pp. 47-52.
11) Yuri Kosarev, Izolda Lee, Andrey Ronzhin, Alexey Karpov, Jesus Savage, Fred Haritatos. Robust Speech Understanding for a Voice Control System. International Workshop SPECOM'2002, St. Petersburg: "Evropeiski Dom", 2002, pp. 13-18.
12) Yuri Kosarev, Izolda Lee, Andrey Ronzhin, Jesus Savage. State of the Art in Speech Understanding. International Workshop SPECOM'2001, Moscow: Moscow State Linguistic University, 2001, pp. 241-250.

Moreover the publications have been submitted and published in the following reviewed journals:

1) Andrey Ronzhin, Yuri Kosarev, Alexey Karpov, Izolda Lee. Recognition and understating of continuous speech based on sliding analysis and integral structure of processing. Scientific journal "Artificial Intelligence" of the Institute of Artificial Intelligence of the National Academy of Sciences of Ukraine, Volume 4, 2003, pp. 430-437.
2) A.L. Ronzhin, Y.A. Kosarev, I.V. Lee, A.A. Karpov. Continuous speech recognition method based on a signal analysis within a sliding window and fuzzy sets theory. Scientific journal "Artificial Intelligence" of the Institute of Artificial Intelligence of the National Academy of Sciences of Ukraine, Volume 4, 2002, pp. 256-263.
3) Yuri Kosarev, Izolda Lee, Andrey Ronzhin and Alexey Karpov. Robust Speech Understanding Methods for New Intellectual Applications, submitted in the International Journal of Speech Technology (ref. # IJST163-03).

Conclusion

The main aim of the project was the development of a voice control model (of a robot, a car, an aircraft, etc.) with a maximal use of data on human speech perception to achieve a maximal accuracy and robustness of control process against different kinds of impeding factors. Voice control is currently considered as one of the most perspective control forms, since it provides maximal efficiency, comfortable work conditions, fast training of personnel, etc. It is especially suitable for extreme, not typical conditions (considerable temperature deviations, mechanical overloads, poor lighting, constant use of hands for steering, etc.).

The main activities of the project were the theoretical (fundamental) research and development of the voice control model including the programming realization. In the framework of the fundamental research the following stages have been fulfilled:

- the analysis of the existent systems for voice control and the methods for speech recognition/understanding;
- selection of the main challenges in the area of speech recognition/understanding and elaboration of the requirements to the voice control model;
- elaboration of effective methods for all levels of speech processing: speech endpoint detection, parametric signal representation, isolated speech recognition, continuous speech recognition, high-level language processing, speech understanding, using situational context, etc.
- elaboration of the mathematical model of speech understanding based on the speech acts theory, the conception of integral processing as well as the results of well-known psycho-physiological experiments concerning human speech perception.

The developed methods have been realized in the research software model of voice control. Thus, this model based on integral speech understanding includes the following methods and modules:

- the method for speech endpoint detection based on spectral entropy analysis;
- two methods for the parametrical signal representation (sign autocorrelation function and spectral-difference features) robust to variations of the signal level and accidental nonlinear spectrum deformations;
- the sliding analysis method for continuous speech recognition robust to grammatical deviations in a pronounced phrase;
- high level processing (associative and situational analysis);
- integral data processing;
- the module for integral adjustment (adaptation) of speech databases.

This software complex has shown good experimental results and at the same time it has provided comfortable speech interface: naturalness of the speech input, accuracy and robustness of the speech understanding process and adaptability of the model to diverse aspects of

exploitation. It can serve for the research of the speech communication process and can serve as the base for the creation of voice control systems for diverse applied areas.

During the Project the developed voice control model was applied for the concrete object domain. The demonstration version of the voice operated flying object has been elaborated. This model includes the demonstration versions of the voice control model and the aircraft emulator. The elaboration of this model has been fulfilled in several stages:

- compiling the initial data for the elaboration of the voice operated flying object;
- elaboration of the situational diagram for aircraft control;
- development of the emulator of the aircraft;
- adaptation of the voice control model to the aircraft control task;
- testing and debugging of the model.

Thus, the developed demonstration model is the laboratory prototype of the voice operated flying object, which can be further used for the elaboration of real models of the control object and voice control system.

The possible further research of the group should be oriented to the creation of the speech technologies for mass using. At this moment the group is investigating the speech data organization based on ontologies and in the future it is planned to use the multiagent system for effective collection and updating of speech and language information. Besides, speech can be successfully used jointly with other kinds of modality (for instance handwriting, gestures, etc) for the creation of multimodal interfaces, which can be used for organization of human-computer interaction in perspective intellectual systems, such as modern network systems, new intelligent services and applications, new generation of mobile devices, etc., where speech becomes the most perspective means for obtaining information.

References

1. A Taxonomy of Multimodal Interaction in the Human Information Processing System, Report of the ESPRIT PROJECT 8579, February 1995.
2. Ajmera J., McCowan I., Bourlard H. Speech/music segmentation using entropy and dynamism features in a HMM classification framework, Speech Communication 40 (2003) 351–363, 2003.
3. Akinori I., Chiori H., Masaharu K., Masaki K. Language Modeling by Stochastic Dependency Grammar for Japanese Speech Recognition. – In Proceedings of ICSLP'2000, Beijing, China, 2000. – pp. 441-444.
4. Allen J., Miller B., Ringger E., Sikorski T. Robust Understanding in a Dialogue System. Proc. ACL, 1996.
5. Austin J. L. How to do things with words. – New York: "Oxford University Press", 1973.
6. Bahl L. R., Jelinek F., Mercer R. A maximum likelihood approach to continuous speech recognition. – IEEE Trans. Pattern Anal. Machine Intell., 1983. – vol. PAMI-5, pp. 179-190.
7. Bahl L. R. et al. Performance if the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task. – In Proceedings of ICASSP, 1995. - pp. 41-44.
8. Ball G., Hall D. ISODATA, A Novel Method of Data Analysis and Patten Classification. - (AD 699616) California, Stanford Research Institute, 1965.
9. Bakis R. Continuous speech word recognition via centisecond acoustic states. – Washington: In Proc. ASA Meeting, 1976.
10. Becker N. Multimodal Interface For mobile clients, Technical report TRITA-NA-E01102, December 2001.
11. Bellegarda J., Silverman K. Toward Unconstrained Command and Control: Data-Driven Semantic Interface. – In Proceedings of ICSLP'2000, Beijing, China, 2000. – pp. 576-579.
12. Bellman R. E. Dynamic programming. – Princeton, New Jersey: Princeton University Press, USA, 1957.
13. Bladon R.A. Problem of normalizing the spectral effects of variations in the fundamental. In Proceedings of the Institute of Acoustics Autumn Conference, 1982.
14. Bonneau-Maynard H., Devillers L. A Framework for Evaluating Contextual Understanding. - In Proceedings of ICSLP'2000, Beijing, China, 2000. – pp. 1734-1737.
15. Carpenter B., Lerner S., Pieraccin R. Optimizing BNF Grammars through Source Transformations. - In Proceedings of ICSLP'2000, Beijing, China, 2000. – pp. 1218-1221.

16. Chien J. On-line Hierarchical Transformation of Hidden Markov Models for Speaker Adaptation. Proc. 1998 ICSLP.
17. Chomsky N. On certain formal properties of grammars. - Inform. Control 2, 1959.
18. Chou W., Zhou Q., Kuo H., Saad A., Attwater D., Durston P., Farrell M., Scahill F. Natural Language Call Steering for Service Applications. – In Proceedings of ICSLP'2000, Beijing, China, 2000.
19. Cognition and the symbolic processes/Edited by Weimer W.. Palermo D. – Hillsdale, 1974.
20. Cohen M., Franco H., Morgan N., Rumbelhart D., Abrash V. Hybrid neural network/Hidden Markov Model continuous speech recognition. Proc. ICSLP, 1992.
21. Cox R.V., Kamm C.A., Rabiner L.R., Schroeter J. and Wilpon J.G. Speech and Language Processing for Next-Millennium Communications Services, Proceedings of the IEEE, Vol. 88, No. 8, pp. 1314-1337, August 2000.
22. Danejko M. Maschkina L., Nechaj O., Sorkina W., Saharanda A. Statiatische Untersuchung der lexikalischen Distribution der Wortformen. - In Sprachstatistik. Mit zahlreichen Skizzen, Tabellen und Schemata im Text. Uebersetzt von einem Kollektiv unter Leitung von Lothar Hoffman. Wilhelm Fink, Muenchen/Salzburg, 1973.
23. Davis S., Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. – In Proceedings of ASSP'28, 1980.
24. Deese J. On the structure of associative meaning. – In Psychological review, 1962. – Vol. 69, No. 2. – pp. 161-175.
25. Esteve Y., Bechet F., R. de Mori. Dynamic Selection of Language Models in a Dialogue System. - In Proceedings of ICSLP'2000 , Beijing, China, 2000.
26. Freeman D., Sonthcott C., Boyd I. A Voice Activity Detector for the Pan-European Digital Cellular Mobile Telephone Service. IEE Colloquium "Digitized Speech Communication via Mobile Radio". 1988. p.p. 6/1-6/5.
27. Fujimoto M., Ariki Y. Evaluation of noisy speech recognition based on noise reduction and acoustic model adaptation on the AURORA2 tasks. Proc. Int. Conf. on Spoken Lang. Processing ICSLP'2002, Denver, USA, 2002.
28. Furui S. and Matsui T. Model-based unsupervised instantaneous speaker adaptation. Proc. Acoustical Society of America 132nd meeting - Hawaii, December 1996.
29. Gangemi A., Pisanelli D.M., Steve G. An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies // Data & Knowledge Engineering, 1999. V. 31. Pp. 183—220.
30. Geutner P., Arevalo L., Breuninger J. VODIS – Voice-Operated Driver Information Systems: A Usability Study on Advanced Speech Technologies for Car Environments. – In Proceedings of ICSLP'2000, Beijing, China, 2000.

31. Gruber T.R. Towards principles for design of ontologies used for knowledge sharing. Technical report, Stanford University, Palo Alto, CA, 1993.
32. Guarino N. Understanding, Building, and Using Ontologies. A Commentary to "Using Explicit Ontologies in KBS Development" (by van Heijst, Schreiber, and Wielinga) // International Journal of Human and Computer Studies, 1997. V. 46. № 2/3. Pp. 293 – 310.
33. Handbook of Human-Computer Interaction, (ed. by J. Jacko & A. Sears), Lawrence Erlbaum: New Jersey, 2002.
34. Hermansky H., Morgan N. RASTA Processing of Speech. IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 4, October 1994. – pp. 578-589.
35. Hirasawa J., Miyzaki N., Nakano M., Aikawa K. New Feature Parameters for Detecting Misunderstanding in Spoken Dialogue System. – In Proceedings of ICSLP'2000, Beijing, China, 2000.
36. Homma S., Takahashi J. and Sagayama S. Iterative Unsupervised Speaker Adaptation for Batch Dictation. Proc. 1996 ICSLP.
37. Homma S., Aikawa K., Sagayama S. Improved Estimation of Supervision in Unsupervised Speaker Adaptation Proc. 1997 ICASP.
38. Horiuchi Y., Arsushi F., Ichikawa A. New WWW Browser for Visually Impaired People Using Interactive Voice Technology. – Budapest, Hungary: In Proc. Of Eurospeech'99, 1999. – pp. 2139-2142.
39. Howes D. On the relation between the probability of a word as an association and in general verbal usage. – In Journal of Abnormal and Social Psychology, 1957. – Vol. 54, No. 1.
40. http://www.elsnet.org
41. http://www.elra.info/
42. Huang Y., Zheng F., Xu M., Yan P., Wu W. Language Understanding Component for Chinese Dialogue System. – In Proceedings of ICSLP'2000, Beijing, China, 2000. – pp. 858-862.
43. Ishii J. Speaker Normalization and Adaptation Based on Linear Transformation. ICASSP'97, Vol. 2, 1997 – pp. 1055-1058.
44. Jelinek F. A fast sequential decoding algorithm using stack. – IBM J. Res. Develop., 1969. – No. 13. – pp. 675-685.
45. Jelinek F. The Development of an Experimental of Discrete Dictation Recognizer. – In Proceedings of IEEE, No. 11, Vol. 73, 1985.
46. Jelinek F. Statistical methods for speech recognition. Massachusetts Institute of Technology, 1999.
47. Johnsen M., Holter T., Svendsen T., Harborg E. Stochastic Modeling of Semantic Content for Use in a Spoken Dialogue System. - In Proceedings of ICSLP'2000, Beijing, China, 2000.

48. Johnson S. C. Hierarchical clustering schemes. Psychometrika. – 1967. – 32.
49. Johnson S. and WoodLand P. Speaker Clustering Using Direct Maximisation of the MLLR-Adapted Likelihood, Proc. 1996 ICSLP.
50. King B. F. Step-wise clustering procedures. - Journal of the American Statistical Association, 1967. – 62.
51. Klatt D.H. Prediction of perceived phonetic distance from critical-band spectra: A first step. In Proceedings of the Int. Conf. Acoust. Speech Signal Processing, 1982.
52. Kosarev Yu. A., Jarov P. A. Associations help to recognize words. – In Proceedings of DAGA-95, Saarbruecken, 1995. – pp. 979-982.
53. Kosarev Yu., Piotrowski R. Synergetics and 'Insight' Strategy for Speech Processing. Literary and Linguistic Computing - Oxford University Press, Vol. 12, № 2, 1997.
54. Kosarev Yu., Savage J. Realization of some reserves of language and extralinguistic knowledge for the speech dialogue systems improvement. Moscow: Proc. Intern. Workshop "Speech and Computer", SPECOM'1999. – pp. 20-31.
55. Kosarev Yu. Some aspects of Robust Speech Understanding. Invited lecture for the International Workshop SPECOM'2002, St. Petersburg, 2002. – pp. 3-8.
56. Kosarev Yu. A., Ronzhin A., Lee I., Karpov A., Savage J., Haritatos F. Robust Speech Understanding for Voice Control System. International Workshop SPECOM'2002, St. Petersburg, 2002. – pp. 13-18.
57. Kosarev Yu., Ronzhin A., Lee I. and Karpov A. "Continuous Speech Recognition without Use of High-Level Information", 15-th International Congress of Phonetic Sciences, August 2003, Barcelona, Spain, pp. 1373-1376.
58. Kosarev Yu., Lee I., Ronzhin A., Karpov A. "Robust Speech Understanding Methods for New Intellectual Applications", AVIOS 2003 Conference, San Jose, California, USA.
59. Kravez L. G. Quantitative Merkmale englischer Nominalverbindungen. - In Sprachstatistik. Mit zahlreichen Skizzen, Tabellen und Schemata im Text. Uebersetzt von einem Kollektiv unter Leitung von Lothar Hoffman. Wilhelm Fink, Muenchen/Salzburg, 1973.
60. Kuo H., Lee C. Discriminative Training in Natiral Language Call Routing. - In Proceedings of ICSLP'2000, Beijing, China, 2000.
61. Kurematsu A., Akegam Y., Burge S., Jekat S., Lause B., Maclaren V., Oppermann D., Schultz T. VERBMOBIL Dialogues: Multifaced Analysis. – In Proceedings of ICSLP'2000, Beijing, China, 2000.
62. Lai Y., Lee K., Wu C. Intention Extraction and Semantic Matching for Internet FAQ Retrieval Using Spoken Language Query. - In Proceedings of ICSLP'2000, Beijing, China, 2000.

63. Leonardi F., Micca G., Militello S., Nigra M. Preliminary results of multilingual interactive voice activated telephone service for people-on-the-move. – In Proceedings of EUROSPEECH'97, 1997. - vol. 4, pp. 1771-1774.
64. Levinson S. E. "Structural Methods in Automatic Speech Recognition." – In Proceedings of the IEEE, 1985. - vol. 73, no. 11, pp. 1625-1650.
65. Lin Y., Wan H. Error-tolerant Language Understanding for Spoken Dialogue Systems. - In Proceedings of ICSLP'2000, Beijing, China, 2000.
66. Lindsay P. H., Norman D. A. Human Information Processing. – NY and London: Academic Press, 1972.
67. Lowerre B., Reddy D. The Harpy speech understanding system. – Pittsburgh: Carnegie – Mellon University, 1976.
68. Lucke H. Interface of stochastic context-free grammar rules from example data using the theory of Bayesian belief. – In: The Proc. of Eurospeech'93, 1993. – pp. 1195-1198.
69. Luo X., Franz M. Semantic Tokenization of Verbalized Numbers in Language Modeling. – In Proceedings of ICSLP'2000, Beijing, China, 2000.
70. Lyons J. Introduction to theoretical linguistics. – Cambridge: At the University Press, 1972.
71. McGregor J.D., Sykes D.A, Practical Guide to Testing Object-Oriented Software, USA:Addison-Wesley, 2001.
72. MacQueen J. B. Some methods for classification and analysis of multivariate observations. – In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. – 1967.
73. Maglio P. P., Matlock T., Campbell C. S., Zhai S. and Smith B. A. Gaze and Speech in Attentive User Interfaces, Proc. of the Third International Conference on Multimodal Interfaces, Beijing, China, 2000.
74. Makhoul J., Raucos S., Gish H. Vector Quantization In Speech Coding. - In Proceedings of IEEE, 1985. – vol. 73, No. 11, pp. 1551-1588.
75. Markel J., Gray A. Linear Prediction of Speech. – Berlin, Springer-Verlag, 1976.
76. Matsui T., Furui S. N-best Based Instantaneous Speaker Adaptation Method for Speech Recognition. Proc. 1996 ICSLP.
77. Miller G., Isard S. Some Perceptual Consequences of Linguistic Rules, J. of Verbal Learning and Verbal Behavior, 1963. - 2, pp. 217-228.
78. Miyazawa Y., Takami J., Sagayama S. and Matsunaga S. All-phoneme Ergodic Hidden Markov Network for Unsupervised Speaker Adaptation. Proc. 1994 ICASSP.
79. Myers C. S., Rabiner L. R. A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition. - IEEE Trans. ASSP-29, 1981. – No. 2, pp. 284-297.
80. Myrvoll T., Siohan O., Lee C., Chou W. Structural Maximum a Posteriori Linear Regression for Unsupervised Speaker Adaptation. – In Proceedings of ICSLP'2000, Beijing, China, 2000. – pp. 78-81.

81. Oaksford M., Chater N. Against logistics cognitive science. – In Mind &Language, 1991. – Vol. 6, No. 1, pp. 2-37.
82. Ono Y., Wakita H., Zhao Y. Speaker Normalization Using Constrained Spectra Shifts in Auditory Filter Domain. Eurospeech'93, Vol.1, 1993. – pp. 355-358.
83. Oviatt, S.L. Multimodal system processing in mobile environments. Proceedings of the Thirteenth Annual ACM Symposium on User Interface Software Technology (UIST'2000), 21-30. New York: ACM Press, 2000.
84. Pickles J. O. An Introduction to the Physiology of Hearing. – New York: Academic press, USA, 1988.
85. Picone J. Continuous Speech Recognition Using Hidden Markov Models. IEEE ASSP Magazine, Vol. 7, No. 3, July 1990.
86. Picone J. Signal Modeling Techniques In Speech Recognition. IEEE Proceedings, Vol. 81, No. 9, 1993.
87. Potamianos A., Kuo H. Statistical Recursive Finite State Machine Parsing for Speech Understanding. - In Proceedings of ICSLP'2000, Beijing, China, 2000. – pp. 1237-1240.
88. Rabiner L. R., Schafer R. W. Digital Processing of Speech Signals. – New Jersey: Prentice-Hall, Englewood Cliffs, USA, 1978.
89. Rabiner L.R., Wilpon J.G. and Juang B.H., "A Model-Based Connected Digit Recognition System Using Either Hidden Markov Models or Templates", Computer Speech and Language, 1 (2): 167-197, December 1986.
90. Rabiner L.R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. IEEE, 77 (2): 257-286, February,1989.
91. Rabiner L., Juang B. Fundamentals of Speech Recognition. – New Jersey: Prentice-Hall, Englewood Cliffs, USA, 1993.
92. Rahim M., Pieaccini R., Eckert W., Levin E., Di Fabbrizio G., Riccardi G., Kamm C., Narayanan S. A Spoken Dialogue System for Conference / Workshop Services. – In Proceedings of ICSLP'2000, Beijing, China, 2000.
93. Ronjin A., Lee I., Kosarev Yu. Quasi-allophone method of acoustic + voice adaptation. Proc. SPECOM 2000. – pp. 91-93.
94. Ronzhin A., Lee I., Kosarev Yu., Karpov A. Continuous Speech Recognition Method Suitable for Robust Speech Understanding. International Workshop SPECOM'2002. St. Petersburg, 2002, pp. 47-52.
95. Ronzhin A., Kosarev Y., Karpov A., Lee I. Elaboration of the intellectual speech interface provided accuracy, robustness and adaptability. International Workshop SPECOM'2003, Moscow, Russia, October 2003.
96. Ronzhin A., Kosarev Y., Karpov A., Lee I. Recognition and understating of continuous speech based on sliding analysis and integral structure of processing. Scientific journal "Artificial Intelligence" of the Institute of Artificial Intelligence of the National Academy of Sciences of Ukraine, Volume 4, 2003, pp. 430-437.

97. Sakoe H., Chiba S. Recognition of Continuously Spoken Words based on Time-Normalization by Dynamic Programming. – J. Acoust. Soc. Japan, 1971 - 7, 9.
98. Sakoe H. Two-Level DP Matching – A Dynamic Programming-Based Pattern Matching Algorithm for Connected Word Recognition. - IEEE Trans. ASSP-27, 1979. – No. 6, pp. 588-595.
99. Selfridge M. Integrated Processing Produces Robust Understanding. – "Computational Linguistics", Vol. 12, №2, April-June 1986. – pp.89-106.
100. Seward A. A Tree-Trellis N-best Decoder for Stochastic Context-Free Grammars. - In Proceedings of ICSLP'2000, Beijing, China, 2000. – pp. 1032-1035.
101. Schank R. Conceptual Information Processing. – Amsterdam, North-Holland, 1975.
102. Schank R., Birnbaum L., May J. Integrating semantics and pragmatics. – «Quaderni di Semantica», 1985. - Vol. VI, no. 2.
103. Schank R.C., Abelson R.P. Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures. Hillsdale (NY), 1997.
104. Shen J.-L., Hung J.-W. and Lee L.-S. Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments, Proc. Int. Conf. on Spoken Lang. Processing ICSLP'98, Sydney, Australia, 1998.
105. Sonthcott C. Speech Proceeding in the Pan-European Cellular Mobile Telephone System. IEE Colloquium "Digitized Speech Communication via Mobile Radio". 1988. pp. 5/1-5/5.
106. Strom N. Continuous Speech Recognition in the WAXHOLM Dialogue System. – STL QPSR, 1996. – pp. 67-95.
107. Suzuki M., Abe T., More H., Makino S. and Aso H. High-Speed Speaker Adaptation Using Phoneme Dependent Tree-Structured Speaker Clustering. Proc. 1998 ICSLP.
108. Swerts M., Litman D., Hirschberg J. Corrections in Spoken Dialogue Systems. – In Proceedings of ICSLP'2000, Beijing, China, 2000.
109. Takahashi J. and Sagayama S. Vector-Field-Smoothed Bayesian Learning for Incremental Speaker Adaptation. Proc. ICASSP'1995.
110. Trends in Speech Recognition. Ed. Lea W.A., Prenrice-Hall, Inc. Englewood Cliffs, New Jersey, 1980.
111. Tuerk C., Robinson T. A new Frequency Shift Function for Reducing Inter-Speaker Variance. Eurospeech'93, Vol.1, 1993. – pp. 351-354.
112. Varile G., Zampolli A. Survey of the State of the Art in Human Language Technology. Cambridge University Press, 1997.
113. Vilar J., Llorens D, Vidal E. Experiments with Finite-State Models for Speech-Input Language Translation. – In Proceedings of SPECOM'96. – St-Petersburg, 1996. – pp. 59-63.
114. Vintsiuk T. K. Element-Wise Recognition of Continuous Speech Consisting of Words from a Specified Vocabulary. – Kibernetika, 1971. – No. 2, pp. 133-143.

115. Viterbi A. J. Error bounds for convolutionalcodes and an asymmetrically optimum decoding algorithm. – IEEE Transactions on Information Theory, 1967. – vol. IT-13, pp. 260-267.
116. Waheed K., Weaver K. and Salam F.M. A robust algorithm for detecting speech segments using an entropy contrast, Proc. 45th IEEE International Midwest Symposium on Circuits and Systems MWSCAS'2002, Oklahoma, USA, 2002.
117. Wang J., Wang H., Lee K., Huang C. Domain-unconstrained language understanding Based on CKIP-Auto Tag, How-net, and ART. – In Proceedings of ICSLP'2000, Beijing, China, 2000. – pp. 807-810.
118. Wang Yu and Zhu Xiaoyan. A New Approach for Incremental Speaker Adaptation, Proc. 2000 ICSLP.
119. Wang H. M., Lin Y. C. Coal-oriented Table-driven Design for Dialog manager. – In Proceedings of ICSLP'2000, Beijing, China, 2000.
120. Wozencraft J., Reiffen. B. Sequential decoding. – Technology Press and Wiley, New York, 1961.
121. Wu C., Chen Y., Yang C. Error Recovery and Sentence Verification Using Statistical Partial Pattern Tree for Conversational Speech. - In Proceedings of ICSLP'2000, Beijing, China, 2000.
122. Yokoo A., Sagisaka Y., Campbell N., Iida H., Yamamoto S. ATR-MATRIX: Speech Translation System from Japanese to English. – In Proceedings of SPECOM'98. – St-Petersburg, 1998. – pp. 203-206.
123. Zadeh L. «A fuzzy-algorithmic approach to the definition of complex or imprecise concepts». In International Journal of Man-Machine Studies. Vol. 8, No. 3, 1976.
124. Zhang H., Xu B., Huang T. How to Choose Training Set for Language Modelling. In Proceedings of ICSLP'2000, Beijing, China, 2000.
125. Zhao Y. Self-Learning Speaker Adaptation Based on Spectral Variation Source Decomposition Eurospeech'93, Vol.1, 1993. – pp. 359-362.
126. Zwicker E., Terhardt E. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. – Journal of the Acoustical Society of America, 1980. – vol. 68, No. 5, pp. 1523-1525.

127. Анализ, распознавание и интерпретация речевых сигналов / Винцюк Т.К. – Киев: Наук. думка, 1987.
128. Вейценбаум Дж. Гл. VII «Понимание связного текста вычислительной машиной». Сб. Распознавание образов. Исследование живых и автоматических распознающих систем. Москва, 1970, с.229.
129. Винцюк Т.К. Куляс А.И. Универсальная программа анализа речи в реальном масштабе времени // 10 Всесоюзный семинар «Автоматическое распознавание слуховых образов»: Тез. докл. – Тбилиси, 1978.

130. Винцюк Т.К. Два основных пути создания систем распознавания и смысловой интерпретации слитной речи // 11 Всесоюзный семинар «Автоматическое распознавание слуховых образов»: Тез. докл. – Ереван, 1980. – С. 221-225.
131. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. – Киев: Наук. думка, 1987. – 264 с.
132. Винцюк Т.К., Скрипник А.Г. Модуль анализатора речи СРД “Речь-2” – Тезисы докладов 16-го всесоюзного семинара (АРСО – 16), 1991. – с. 250-251.
133. Галунов В.И., Галунов Г.В. Вариант системы распознавания речи. Труды Международного семинара Диалог'2000, т.2, Протвино, 2000.
134. Гойхман О.Я., Надеина Т.М. Речевая коммуникация. М., Инфра – М, 2003, 10 с.
135. Дрейфус Х. Чего не могут вычислительные машины. Пер. с англ. – М.: Прогресс, 1978. – 336 с.
136. Ершов А.П. К методологии построения диалоговых систем: феномен деловой прозы // Вопросы кибернетики: Общение с ЭВМ на естественном языке. – М.: Наука, 1982. – Вып. 80. – С. 3-20.
137. Иванова Т.И. Компьютерные технологии в телефонии. Эко-Трендз, М., 2002.
138. Инструкция по взаимодействию и технология работы членов экипажа самолета Ил-76 (Ил-76Т), Москва “Воздушный транспорт”, 1984.
139. Карпов А.А., Косарев Ю.А., Ронжин А.Л., Ли И.В. Система акустических признаков речевого сигнала, устойчивых к вариациям уровня громкости и спектра сигнала. Труды III Всероссийской конференции «Теория и практика речевых исследований» АРСО-2003. Москва, МГУ им. М.В. Ломоносова, Сентябрь 2003г., с.83-88.
140. Кельманов А.В. О некоторых проблемах построения систем распознавания инвариантных к диктору // 15 Всесоюзный семинар «Автоматическое распознавание слуховых образов»: Тез. докл. – Таллинн, 1989. – С. 103-104.
141. Классификация и кластер / Под ред. Райзина Дж.В. – М.: Мир, 1980 – 389 с.
142. Коган А. Состав личности и некоторые ее свойства в контексте ситуационного анализа..// Сборник трудов Института психологии им.С.Г. Костюка АПН Украины – Том 2, часть 3 – Киев, 2000. – С.80-92.
143. Косарев Ю.А. Естественная форма диалога с ЭВМ. – Л.: Машиностроение, 1989. – 143 с.
144. Косарев Ю.А., Ли И.В., Ронжин А.Л., Savage J. Методы понимания речи и текста. Труды СПИИРАН / Под ред. Р.М. Юсупова вып. 1, Т. 2 – СПб.: «Анатолия», 2002. – С. 157-195.
145. Крестьянинов С.В. Интеллектуальные сети и компьютерная телефония. М., «Радио и связь», 2001.
146. Леонтьев А.А. Основы психолингвистики. М.,Смысл., 2003, 276 с.

147. Мазуренко И. Л., Многоканальная система распознавания речи, Сборник трудов VI всероссийской конференции "Нейрокомпьютеры и их применение", Москва, 2000 г.
148. Мельчук И.А. Опыт теории лингвистических моделей «Смысл-Текст». Москва, Наука, 1974.
149. Моттль В.В., Мучник И.Б. Скрытые марковские модели в структурном анализе сигналов / М.: Физматлит, 1999.-351с.
150. Мясников Л.Л. Объективное распознавание звуков речи // ЖТФ. – 1943. –№ 3. – С. 109-115.
151. Наставление по производству полетов в гражданской авиации, Москва «Воздушный транспорт», 1985.
152. Пиотровский Р.Г. Текст, машина, человек. – Л.: Наука, 1975. – 327 с.
153. Покровский Н.Б. Расчет и измерение разборчивости речи. – М.: Связь, 1962. – 391 с.
154. Поспелов Д.А. Ситуационное управление. Теория и практика. М. Наука 1986.
155. Распознавание слуховых образов. / Под ред. Загоруйко Н.Г. – Новосибирск: «Наука», 1970. – 340 с.
156. Ронжин А.Л., Косарев Ю.А., Карпов А.А., Ли И.В. Распознавание и понимание слитной речи на основе скользящего анализа и интегральной структуры обработки. Научно-теоретический журнал «Искусственный интеллект». №4 – Донецк, Украина, 2003, с.430-437.
157. Руководства по летной эксплуатации самолета Ил-76Т(ТД), Москва “Воздушный транспорт”, 1984.
158. Сапожков М.А. Речевой сигнал в кибернетике и связи. – М.: Связьиздат, 1963. – 452 с.
159. Сорокин В.Н. Теория речеобразования. – М.: Радио и связь, 1985.
160. Трунин-Донской В.Н. Опознавание набора слов с помощью цифровой вычислительной машины. // Работы по технической кибернетике. – М.: ВЦ АН СССР, 1967. – С. 37-51.
161. Ушакова Т.Н. Проблема внутренней речи в психологии и психофизиологии // Психологические и психофизиологические исследования речи. – М.: Наука, 1985. – С. 13-26.
162. Фант. Г. Акустическая теория речеобразования. Пер. с англ. – М.: Наука, 1964. – 284 с.
163. Шабес В.Я. Речь и знание. СПб., 1992.